

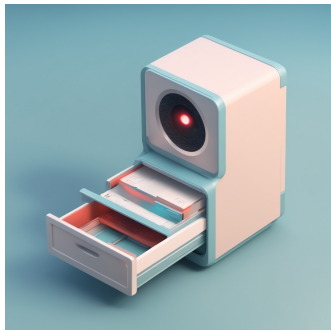
AI assistant for information retrieval

Alexandr Prozorov

April 24, 2025

AI in Science: Reflection, Opportunity & The Road Ahead

- Dispersed collaboration documents makes it hard to have a **comprehensive** overview
- Manual search is not comprehensive, correlation between distinct documents
- Plots search, debugging and code snippets
- Data preservation tool
- The **chatbot** enables:
 - Quick **access** to relevant papers, code, and documentation.
 - Faster **onboarding** for new collaborators.
 - Most of the data is **private** (no google/chatgpt)
 - **Coding** assistance



AI-Powered Research Assistant

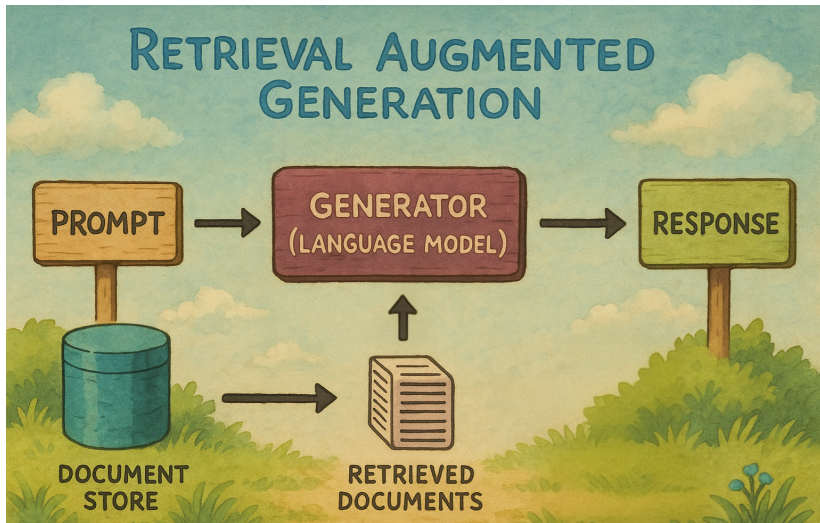
Why LLMs in Science?

- LLMs "read" many documents in **seconds**.
- They **contextualize** results (with retrieval)
- Biggest asset is written knowledge — papers, logs, chats, codes

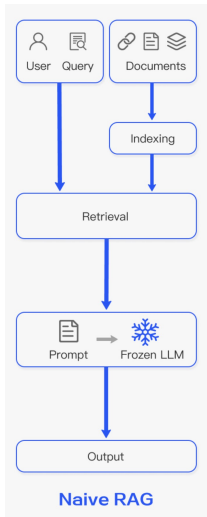
Isn't AI Enough? Why Still Use RAG?

- LLaMA4 (Apr 25) has **1M** context.
- But: big experiments have diverse, noisy, evolving data.
- Different techniques in AI - fine-tuning, RLHF, etc. - are **expensive** and **time-consuming**.
- RAG ensures precision, **traceability**, and trust by grounding LLMs in curated, real **sources**.

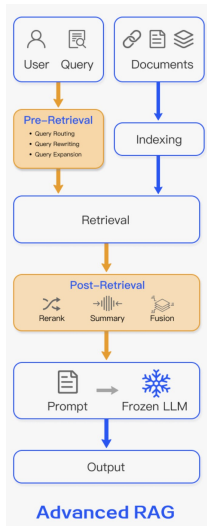
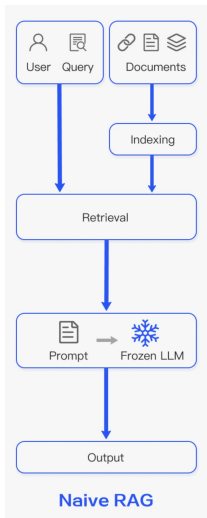




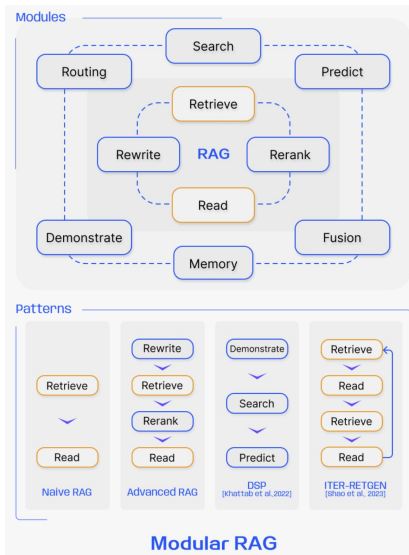
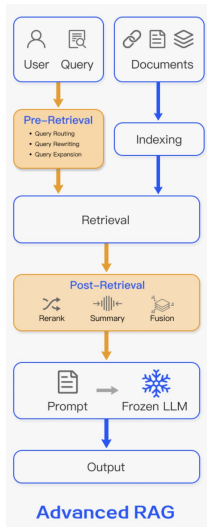
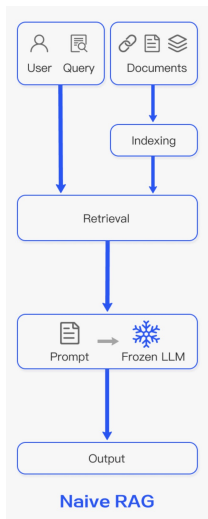
RAG paradigms













RAG paradigms



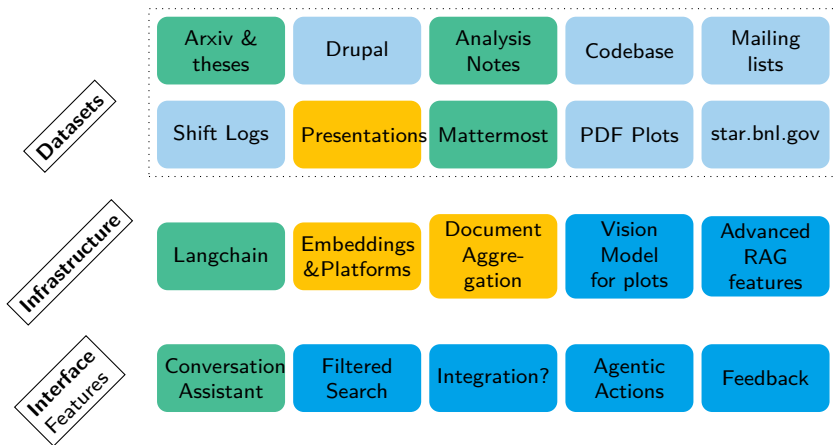
RAG paradigms



Techologies

- Langchain  - orchestrator, core **logic**
- **Embeddings** - SentenceTransformer 
- **Database**  Chroma  Pinecone
- **Interface**  Streamlit will be changed later to sophisticated
- **Data extraction**   
- **Web Scrappers**  

Current Plans for STAR experiment as example



Usage

- ePIC - [rags4eic](#)
- Searching BNL / SDCC Web pages
- Assists users in trivial issues (if public, information could be correlated)
- Global search like a library search of all publications not only a single experiment

From Prototype to Collaboration-Wide Tool

- **Already integrated:**

- Arxiv, theses, starnotes, Mattermost.
- Modular architecture for expanding to new data.
- First fully **local** version (inc. database and LLM) at BNL

- **Planned expansion:**

- Meeting records, mailing lists, shift logs, codebase.
- Specialized agents for targeted search (papers, plots, comments, connections).
- Advanced RAG:
 - Real-time search and retrieval
 - Reranking retrieved sources
 - Iterative retrieval
- Fine-tuning local LLM

From Prototype to Collaboration-Wide Tool

- **Already integrated:**
 - Arxiv, theses, starnotes, Mattermost.
 - Modular architecture for expanding to new data.
 - First fully **local** version (inc. database and LLM) at BNL
- **Planned expansion:**
 - Meeting records, mailing lists, shift logs, codebase.
 - Specialized agents for targeted search (papers, plots, comments, connections).
 - Advanced RAG:
 - Real-time search and retrieval
 - Reranking retrieved sources
 - Iterative retrieval
 - Fine-tuning local LLM

Real-life example

Choose Database

☒ Pinecone (Cloud)

☐ Chroma (Local)

Number of chunks to retrieve

100

10 200

Data search type:

Everything



How do I connect to the SDCC?



Here are the main steps to connect to the SDCC (Scientific Data and Computing Center):

- **Initiate a connection:** Start by connecting to the SDCC using the SSH protocol. Use the command `ssh -X your_username@ssh.sdcc.bnl.gov`, where "your_username" is your SDCC account name¹.
- **Agent start-up:** Before connecting, ensure you have started an agent where you are working².
- **Terminal access:** After connecting, use the command `rterm` to open a terminal from one of the interactive nodes³.
- **Authentication:** You will need your SDCC password, and if you have Multi-Factor Authentication (MFA) enabled, a valid code from your MFA application will be required at each login⁴.
- **Help and troubleshooting:** If experiencing difficulties, you can file a ticket with SDCC support for assistance⁵.

Feel free to reach out for more specific questions!

Footnotes

1. <https://chat.sdcc.bnl.gov/star/pl/n668sqgsbfgxz8o158x3t7mzar> ↗
2. <https://chat.sdcc.bnl.gov/star/pl/unnz9yxklfra38sbn5j9qnrh> ↗
3. <https://chat.sdcc.bnl.gov/star/pl/o67dunfj17rcxrwmiotdx9pcw> ↗
4. <https://chat.sdcc.bnl.gov/star/pl/efhrtmdnepy3bczcotha5u9cnc> ↗
5. <https://chat.sdcc.bnl.gov/star/pl/akqpi8bogbbojpuiinqeqeq16wh> ↗

Current Drupal

Search

Users Content

Enter your keywords:

unfolding in jets

Search

Advanced search

Search results

pAu centrality

... Yanfang Liu a) Presented latest Jpsi RpAu **in** different centrality bins, where the central events are seen to be ... attempts to obtain the true multiplicity through **unfolding** have not been successfully so far due to unstable results ...

Blog entry - marr - 2022-03-02 12:02 - 0 comments - 7 attachments - 1 group

500 GeV Fragmentation Function Analysis Plan

... for underlying event contributions Correct pion-**in**-jet spectrum for underlying event contributions Correct full ... to a pion-**in**-jet at the particle-jet level Step 2: **Unfolding** Unfold full jet p T spectrum, regardless of pions ...

Blog entry - drach09 - 2019-07-01 13:18 - 0 comments - 1 attachment - 0 groups

Update 07.09.2018 -- Run 9 data: 1st stab at unfolding R = 0.3 charged pi0 recoil jets

... corrected reconstruction efficiency and response matrix (**in** the link below), I tried **unfolding** the R = 0.3 charged recoil jet (pi0 trigger) data. ... Also, the labels are incorrect: these are R = 0.3 **jets**. While the results are definitely smoother than the last ...

Blog entry - dimawxc - 2019-02-19 10:49 - 0 comments - 0 attachments - 1 group

Weekly Jet Discussion Meeting

... 2020-06-01 Zilong Chang, combining triggers for **unfolding** Nick Lukow, run15 dijet A_LL update Dmitry Kalinkin, Jet QA with respect to sector 20 **in** run12pp200 2020-04-20 Maria Zurek, Embedding QA - ...

Book page - zchang - 2020-10-04 11:56 - 0 comments - 5 attachments - 1 group

STAR PWGC meeting

... Measurement of Jpsi production within charged **jets** in p+p collisions at sqrt(s) = 500 GeV with the STAR experiment - Brief ... finer bins? If not, can you use finer bins at least **in** the **unfolding**? They can be combined later. There may be certain difference **in** the ...

Meeting - marr - 2021-01-22 12:28 - 1 comment - 0 attachments - 0 groups

Update 09.17.2018 -- Run 9 pp: Unfolding Systematics, Pi0 (R = 0.2, 9 - 11 and 11 - 15 GeV)

... indicate the different systematic variations and the **unfolding** method, regularization parameter, and prior used **in** **unfolding** each variation. Note that the Pythia6 prior is triggered on ...

Blog entry - dimawxc - 2019-02-13 15:53 - 0 comments - 1 attachment - 1 group

Update [01.09.2020] -- Unfolding Run 9 pp Data Using Response Matrices Made With Different DCA Cuts

... feed-down by comparing the track pT distributions **in** the Run 9 dijet embedding sample using different DCA cuts (1 cm vs. 3 ... efficiencies: one where the reconstructed (detector-level) **jets** were made from tracks with DCA < 1 cm, and another



what is unfolding in jets?



- **Unfolding in jets** refers to a set of statistical techniques aimed at correcting for detector effects that alter the measured distributions of jet observables.
- The main purpose is to recover the **true underlying distributions** that would be observed if no detector effects were present, which are essential for accurate physics measurements¹.
- It involves inverting the convolution of spectra and resolution, making it a **non-trivial mathematical problem**².
- Common methods used in unfolding include **Bayesian unfolding**, which employs response matrices constructed from Monte Carlo (MC) simulations to rectify the measured jet spectrum by applying probability theory³.
- **Regularized unfolding** is often employed to handle finite momentum resolution and reconstruction efficiency, allowing corrections to jet spectra measured at the detector level⁴.
- The quality of the unfolding can be tested through methods like **MC closure** and **refold tests**, comparing unfolded results against known distributions⁵.

Footnotes

1. <https://drupal.star.bnl.gov/STAR/starnotes/private/psn0804-0>
2. <https://drupal.star.bnl.gov/STAR/starnotes/private/psn0718>
3. <https://drupal.star.bnl.gov/STAR/starnotes/private/psn0772>
4. <https://drupal.star.bnl.gov/STAR/starnotes/private/psn0804-0>
5. <https://drupal.star.bnl.gov/STAR/starnotes/private/psn0718>

What is under the carpet?

Personal > Tracing projects > default

default

ID Data Retention 14d Add Rule

Runs Threads Monitor Setup

1 filter Last 7 days Root Runs LLM Calls All Runs Columns

		Input	Output	Start Time	Latency	Tokens	Cost
		give me a list of sources containing D0 meson recon...	ai: How can I help ...	23/10/2024, 11:32:25	6.61s	12,294	\$0.00200475
		give me a list of sources containing D0 meson recon...	ai: Hello, here are s...	23/10/2024, 11:31:53	5.70s	12,919	\$0.00210435
		give me a list of sources containing D0 meson recon...	ai: How can I help ...	23/10/2024, 11:31:17	7.71s	12,026	\$0.0019731
		matching procedure for jet unfolding	ai: Here are some k...	23/10/2024, 10:23:29	8.90s	12,176	\$0.00202395
		for jets unfolding	ai: Here are some k...	23/10/2024, 10:22:39	6.77s	11,725	\$0.00194955
		Dmitry Kalinkin jet matching procedure	ai: Hmm, I'm not su...	23/10/2024, 10:21:50	5.96s	9,647	\$0.00159285
		can you give me Dmitry Kalinkin thesis?	ai: How can I help ...	23/10/2024, 10:21:28	3.72s	9,052	\$0.00136635
		Is FTPC still part of STAR? Or was it only for some ru...	ai: Here are some k...	22/10/2024, 10:04:19	6.23s	13,072	\$0.0020868
		Does STAR have any subsystems or techniques to m...	ai: Hmm, I'm not su...	22/10/2024, 10:02:36	8.60s	12,077	\$0.00198525
		Is there nothing at forward rapidity to measure multi...	ai: Here are some k...	22/10/2024, 10:02:06	2.95s	11,483	\$0.0017337

Stats

Last 7 days

RUN COUNT
20

TOTAL TOKENS
223,060 / \$0.04

MEDIAN TOKENS
11,315

ERROR RATE
0%

% STREAMING
100%

LATENCY
P50: 6.13s P99: 8.84s

FIRST TOKEN
P50: 0.46s P99: 1.03s

Filter Shortcuts

Input Key

☐ Input

Input Key Value

☐ Input == *Give Me A Lis...

☐ Input == *Why Proton Is...

What is under the carpet?

TRACE

Collapse Stats Filter Show All ▾

RunnableWithMessageHistory ✓

6.61s 12,294

insert_history 0.00s

RunnableParallel<chat_history> 0.00s

load_history 0.00s

check_sync_or_async 6.61s

retrieval_chain 6.60s

RunnableAssign<context> 1.12s

RunnableParallel<context> 1.12s

retrieve_documents 1.12s

RunnableLambda 0.00s

RunnableSequence 1.12s

ChatPromptTemplate 0.00s

ChatOpenAI gpt-4o-mini 0.60s

StrOutputParser 0.00s

VectorStoreRetriever 0.51s

RunnableAssign<answer> 5.47s

RunnableParallel<answer> 5.47s

stuff_documents_chain 5.47s

format_inputs 0.01s

RunnableParallel<context> 0.01s

format_docs 0.00s

ChatPromptTemplate 0.00s

ChatOpenAI gpt-4o-mini 5.45s



ChatOpenAI

Run Feedback Metadata

Input ▾

SYSTEM

You are an expert on the STAR experiment, a high-energy nuclear physics experiment at Brookhaven National Laboratory. Your task is to answer questions specifically related to technologies, and related topics. Refrain from answering any other topics by saying you will not answer here. DO NOT PROCEED. You are not allowed to use any other sources other than the provided search results (urls and content) and chat history. You must only use information from the provided search results. Generate a comprehensive, and informative answer strictly within 200 words or less provided search results (urls and content) and chat history. You must only use information from the provided search results. Use an unbiased and journalistic tone. Combine search results together into use bullet points in your answer for readability. Make sure to break down your answer into bullet points, do not use any text within <url> and </url> except when citing context. Be specific to the exact questions. Take your time.

Here is the response template:

Response template

- Use bullet points to list the main points or facts that answer the query using the information provided.
- After answering, analyze the respective source links provided within <url> and </url> step. Try to minimize the total number of unique links with no more than 10 unique links.
- You will strictly use no more than 10 most unique links for the answer.
- Use bulleted list of superscript numbers within square brackets to cite the sources correspond to the order of the sources which will be provided in the end of this response.
- End with a closing remark and a list of sources with their respective URLs and title enclosed in the tag <url> and </url> respectively.---

AI

How can I help you?

HUMAN

give me a list of sources containing D0 meson reconstruction for 2016 data AuAu

Future Vision

- Build a science where every decision, insight, and discovery is searchable and preserved.
- AI becomes not just a tool, but a collaborator.
- Fully open-source solutions

Tentative Timeline

- Proof of principle and comprehensive search to STAR, summer 2025
- Extend to other experiment by providing indexing tools, end of 2025
- Finetuning of local LLM, first half 2026
- Unified open-source tool for other science facilities, 2027

Links for CERN and EIC projects:

- [Chatlas Overview Nov. 2023](#)
- [chATLAS IML Meeting Apr. 2024](#)
- [EIC RAG Meeting Jul. 2024](#)

SciBot



Challenges

- **Data Collection**

- Extract source code from web domains and other specific tools.
- - Time-intensive: The scraper must be tailored for each webpage domain.
 - Maintenance: The database requires periodic updates.
- **Advantage:**
 - Knowledge is disentangled from the LLM, eliminating the need for model retraining during database updates.

- **Data Quality**

- Ensure documentation is detailed and up-to-date.
- Avoid duplicates to prevent model inefficiencies.
- Consider privacy



Garbage in - garbage out

Challenges

- **Data Collection**

- Extract source code from web domains and other specific tools.
- - Time-intensive: The scraper must be tailored for each webpage domain.
 - Maintenance: The database requires periodic updates.
- **Advantage:**
 - Knowledge is disentangled from the LLM, eliminating the need for model retraining during database updates.

- **Data Quality**

- Ensure documentation is detailed and up-to-date.
- Avoid duplicates to prevent model inefficiencies.
- Consider privacy



Garbage in - garbage out

Hardware Requirements

Current setup:

- Current LLM: LLaMA v3.3 70 B Inference on one multiple Nvidia GPUs
- or use of OpenAI **API** *gpt4o-mini*
- RAM: 64GB
- Inference time: 4-10 s, depending on input and output text length (<1s for embedding retrieval)
- Database: ChromaDB 5 GB maximum
- or Pinecone **API**

Retrieval

1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4

Production and energy loss of strange and heavy quarks

Lijian Han (for the STAR Collaboration)
Physics Department, Brookhaven National Laboratory, Upton, New York, NY 11973, USA
E-mail: hanlj@prcf.rns.bnl.gov

Abstract
The physics goal at RHIC is to study and identify the properties of matter with partonic degrees of freedom [1]. Hard probes such as identified particles at high p_T from these and jets are thought to be well calibrated probes since they are calculable in the perturbative Quantum Chromodynamics (QCD) framework [2]. The single inclusive hadron production in pp collisions at $\sqrt{s} = 200$ GeV is described by the convolution of parton distribution functions, parton-parton interaction cross sections and fragmentation functions (FFs). In this talk, inclusive particle production at $p_T \sim 4$ GeV can be described by the fragmentation process with jet energy loss. When jet energy loss is taken into account, the jet and those particles that lie away from the jet core and are collimated with increasing p_T [3,4]. This talk is a review of the status of the field. The effects on the particle spectra and the multiplicities from the energy loss are discussed. The effects on the particle spectra and the multiplicities from the energy loss are discussed. The effects on the particle spectra and the multiplicities from the energy loss are discussed.

1. Introduction
The physics goal at RHIC is to study and identify the properties of matter with partonic degrees of freedom [1]. Hard probes such as identified particles at high p_T from these and jets are thought to be well calibrated probes since they are calculable in the perturbative Quantum Chromodynamics (QCD) framework [2]. The single inclusive hadron production in pp collisions at $\sqrt{s} = 200$ GeV is described by the convolution of parton distribution functions, parton-parton interaction cross sections and fragmentation functions (FFs). In this talk, inclusive particle production at $p_T \sim 4$ GeV can be described by the fragmentation process with jet energy loss. When jet energy loss is taken into account, the jet and those particles that lie away from the jet core and are collimated with increasing p_T [3,4]. This talk is a review of the status of the field. The effects on the particle spectra and the multiplicities from the energy loss are discussed. The effects on the particle spectra and the multiplicities from the energy loss are discussed. The effects on the particle spectra and the multiplicities from the energy loss are discussed.

arXiv:1801.1347v1 [nucl-ex] 19 Jan 2018
Production and energy loss of strange and heavy quarks
Lijian Han (for the STAR Collaboration)
Physics Department, Brookhaven National Laboratory, Upton, New York, NY 11973, USA
E-mail: hanlj@prcf.rns.bnl.gov

Abstract
Data taken over the last several years have demonstrated that RHIC has created a hot, dense medium with partonic degrees of freedom. Identified particle spectra at high transverse momentum (p_T) and heavy flavor are thought to be well calibrated probes that serve as ideal tools to study the properties of the medium. We present p_T distributions of particle ratios in pp collisions from STAR experiments to understand the particle production mechanism. These measurements will also constrain fragmentation functions in hadron-hadron collisions. In heavy ion collisions, we highlight: 1) recent measurements of strange hadrons and heavy flavor decay electrons up to high p_T to study jet interaction with the medium and explore partonic energy loss mechanisms; and 2) v_2 and high p_T measurements to study the effect of color screening and other possible production mechanisms.

1. Introduction
The physics goal at RHIC is to study and identify the properties of matter with partonic degrees of freedom [1]. Hard probes such as identified particles at high p_T from heavy flavor and jets are thought to be well calibrated probes since they are calculable in the perturbative Quantum Chromodynamics (QCD) framework [2]. The single inclusive hadron production in pp collisions at $\sqrt{s} = 200$ GeV is described by the convolution of parton distribution functions, parton-parton interaction cross

Retrieval

1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4

arXiv:1901.3347v1 [nucl-ex] 19 Jan 2019

Production and energy loss of strange and heavy quarks

Lijuan Han (For the STAR Collaboration)

Physics Department, Brookhaven National Laboratory, Upton, New York, NY 11973, USA

E-mail: huanli@rcf.rhic.bnl.gov

Abstract

Data taken over the last several years have demonstrated that RHIC has created a hot, dense medium with partonic degrees of freedom. Identified particle spectra at high transverse momentum (p_T) and heavy flavor are thought to be well calibrated probes thus serve as ideal tools to study the properties of the medium. We present p_T distributions of particle ratios in p+p collisions from STAR experiment to understand the particle production mechanisms. These measurements will also constrain fragmentation functions in hadron-hadron collisions. In heavy ion collisions, we highlight: 1) recent measurements of strange hadrons and heavy flavor decay electrons up to high p_T to study jet interaction with the medium and explore partonic energy loss mechanisms; and 2) V and high p_T π^0 measurements to study the effect of color screening and other possible production mechanisms.

1. Introduction

The physics goal at RHIC is to study and identify the properties of matter with partonic degrees of freedom [1]. Hard probes such as identified particles at high p_T or heavy flavor and jets are thought to be well calibrated probes since they are calculable in the perturbative Quantum Chromodynamics (pQCD) framework [2]. The single inclusive hadron production in p+p collisions at $\sqrt{s} = 2$ GeV/c is described by the convolution of parton distribution functions, parton-parton interaction cross

Data taken over the last several years have demonstrated that RHIC has created a hot, dense medium with partonic degrees of freedom.

Identified particle spectra at high transverse momentum (p_T) and heavy flavor are thought to be well calibrated probes thus serve as ideal tools to study the properties of the medium. We present p_T distributions of particle ratios in p+p collisions from STAR experiment to understand the particle production mechanisms.

These measurements will also constrain fragmentation functions in hadron-hadron collisions. In heavy ion collisions, we highlight: 1) recent measurements of strange hadrons and heavy flavor decay electrons up to high p_T to study jet interaction with the medium and explore partonic energy loss mechanisms; and 2) V and high p_T π^0 measurements to study the effect of color screening and other possible production mechanisms.

Retrieval

1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4

ID	VALUES
cf051293-d8e...	-0.0104965866, 0.00714074029, -0.0402951948, 0.0380709899, 0.0462308...

METADATA

arxiv_id: "0812.4157v1"

published: "2008-12-22"

start_index: 3634

text: "with the system size. Even the p+p data yields reasonable fit with the thermal model with a Temperature equi"

title: "Test of Chemical freeze-out at RHIC"

ID	VALUES
d8f83b88-21f...	0.0152690168, -0.117546648, -0.0296285804, 0.0185704585, 0.0439337045,...

METADATA

arxiv_id: "nucl-ex/0412015v1"

published: "2004-12-08"

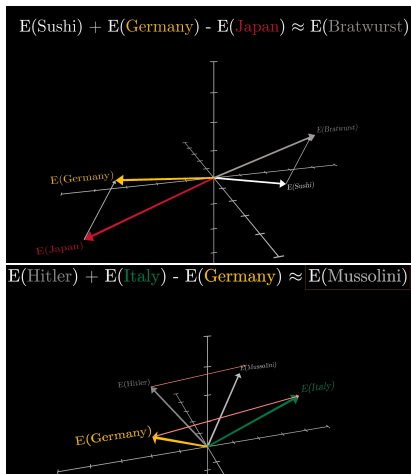
start_index: 735

text: "saturation factor, \sqrt{s} . Another important issue is that statistical models generally utilize Grand Ca"

title: "Volume Effects on Strangeness Production"

Retrieval

1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4



Retrieval

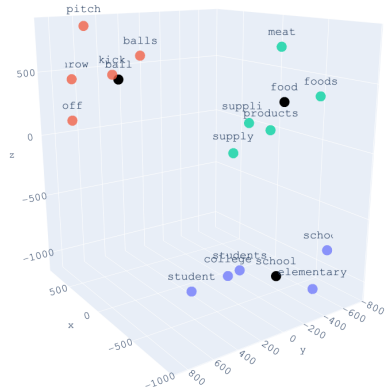
1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4

What is STAR TPC efficiency?



Retrieval

1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4



Retrieval

1. Convert to text
2. Split into K -word chunks
3. Embed each chunk with SentenceTransformer model into 384-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search
6. Append all matching sections to prompt and query GPT-4

Input ▾

1	question: what is TPC efficiency?
YAML ⌵	

DOCUMENTS documents 145

- Due to the change in operation condition, for example, temperature and air pressure of TPC, the
- TPC Acceptance and Tracking Efficiency - ϵ_{TPC} (ϵ_{TPC} is the correction factor)
- TPC tracks. This efficiency is calculated from the STAR data as the ratio of the number of tracks
- same reconstruction chain as the real data production. The TPC efficiency is then calculated as a
- energy loss inside the TPC.
- tracking efficiency depends sensitively on the gain in the proportional readout chamber of the TPC
- measured in the TPC, depends on the momentum. The efficiency of this cut was measured as a function of
- the TPC is high and efforts to understand and quantify the tracking efficiency are underway. The
- The uncertainty in the TPC tracking efficiency is estimated in the same way as in the μ^+ and μ^- channels.

Embeddings and retrieval

Using: [sentence-transformers/all-MiniLM-L6-v2](#)

Model Name	 Performance Sentence Embeddings (14 Datasets) 	Performance Semantic Search (6 Datasets) 	Avg. Performance 	Speed 	Model Size 
sentence-t5-xxl 	70.88	54.40	62.64	50	9230 MB
gtr-t5-xxl 	70.73	55.76	63.25	50	9230 MB
all-roberta-large-v1 	70.23	53.05	61.64	800	1360 MB
all-mpnet-base-v1 	69.98	54.69	62.34	2800	420 MB
gtr-t5-large 	69.90	54.85	62.38	800	640 MB
gtr-t5-xl 	69.88	55.88	62.88	230	2370 MB
all-mpnet-base-v2 	69.57	57.02	63.30	2800	420 MB
sentence-t5-xl 	69.23	51.19	60.21	230	2370 MB
all-MiniLM-L12-v1 	68.83	50.78	59.80	7500	120 MB
sentence-t5-large 	68.74	49.05	58.89	800	640 MB
all-distilroberta-v1 	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v2 	68.70	50.82	59.76	7500	120 MB
all-MiniLM-L6-v2 	68.06	49.54	58.80	14200	80 MB

Embeddings and retrieval

Using: [sentence-transformers/all-MiniLM-L6-v2](#)

- Compact (384-dimensional), free, runs on a local CPU
- Different models on the market
- Multiple “chunks” from one document
- Insert a chunk into database, with metadata of file URL, name, etc.

Search Implemented a child-parent structure for local Chroma database:

- Parent is full doc. Children: Chunks of 200 tokens
- Search children, parent to LLM

Contextualize (with history) user query for efficient retrieval
'history_aware_retriever'

Example of ChatGPT query

Instructions

You are an expert on the STAR experiment ...

Here's a template on how to respond ... Use only information given in the context ...

Retrieved Context (from database)

<context>

...

Title: Observation of D^0 and \overline{D}^0 directed flow in AuAu collisions $s_{NN} = 200\text{GeV}$

type: starnotes

<**url**> drupal.star.bnl.gov/STAR/starnotes/private/psn0715 <**url**/>

Text: presents comparison of p_T - differential $D^0 v_2$ in 0-10% and 10-40% central Au+Au collisions. The $v_2(p_T)$ for D^0 mesons in Run-2014 and Run-2016 are consistent.

...

<context/>

Question (from rephrased prompt)

give me a list of sources containing D0 meson reconstruction for 2016 data AuAu

Example of ChatGPT query

Instructions

You are an expert on the STAR experiment ...

Here's a template on how to respond ... Use only information given in the context ...

Retrieved Context (from database)

<context>

...

Title: Observation of D^0 and \overline{D}^0 directed flow in AuAu collisions $s_{NN} = 200\text{GeV}$

type: starnotes

<**url**>drupal.star.bnl.gov/STAR/starnotes/private/psn0715<**url**>

Text: presents comparison of p_T - differential $D^0 v_2$ in 0-10% and 10-40% central Au+Au collisions. The $v_2(p_T)$ for D^0 mesons in Run-2014 and Run-2016 are consistent.

...

<context/>

Question (from rephrased prompt)

give me a list of sources containing D0 meson reconstruction for 2016 data AuAu

Example of ChatGPT query

Instructions

You are an expert on the STAR experiment ...

Here's a template on how to respond ... Use only information given in the context ...

Retrieved Context (from database)

<context>

...

Title: Observation of D^0 and \overline{D}^0 directed flow in AuAu collisions $s_{NN} = 200\text{GeV}$

type: starnotes

<**url**>drupal.star.bnl.gov/STAR/starnotes/private/psn0715<**url**>

Text: presents comparison of p_T - differential $D^0 v_2$ in 0-10% and 10-40% central Au+Au collisions. The $v_2(p_T)$ for D^0 mesons in Run-2014 and Run-2016 are consistent.

...

<context/>

Question (from rephrased prompt)

give me a list of sources containing D0 meson reconstruction for 2016 data AuAu

What is under the carpet?

Title:Erratum: Observation of D^0 meson nuclear modifications in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV

type:starnotes <url><https://drupal.star.bnl.gov/STAR/starnotes/private/psn0694-0></url>

Text:We have identified two errors in the estimation of the particle identification efficiency. Considering the high combinatorial background in (D^0) meson reconstruction in Au+Au collisions, a hybrid

Title:STAR Collaboration Beam Use Request for run 18 and run 19

type:starnotes <url><https://drupal.star.bnl.gov/STAR/starnotes/public/sn0670></url>

Text:With the Run 14 data, STAR has obtained the first result on (Λ_c) baryon production and significantly improved measurement of (D_s) meson production in Au+Au collisions, shown in Fig.

Title:Results from Large Scale STAR Raw Data Reconstruction on NERSC HPC

type:starnotes <url><https://drupal.star.bnl.gov/STAR/starnotes/public/sn0688></url>

Text:Computing (HTC) workloads. The allocation was successfully applied to reconstruction of the entire 2016 STAR raw data for Au+Au collisions at 200 GeV. This report presents the results and

Title:Observation of D^0 and $\overline{D^0}$ directed flow in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV

type:starnotes <url><https://drupal.star.bnl.gov/STAR/starnotes/private/psn0715></url>

Text:presents comparison of (p_T) -differential $(D^0)(v_2)$ in 0-10% and 10-40% central Au+Au collisions. The $(v_2)(p_T)$ for (D^0) mesons in Run-2014 and Run-2016 are consistent. These results