AI agents in LHCb

Pawel Kopciewicz (CERN)

Introduction

Thank you for inviting me!

- I’m an applied fellow in CERN LHCb DP section

I was wondering how to make this presentation the most interesting, and

I decided to focus on the following:

- Status of the LHCb Brain project
- General considerations on AI in LHCb & selected developments
- AI tools for coding

LHCb Brain

This project attempts to develop and apply frontier AI in LHCb

(Please treat the below as unofficial information yet)

Selected as one of 4 strategic AI initiatives at CERN

- Received funding for AI industry partnership for a 3-month Proof of Concept
- Industry partner ranked us #1 in terms of interest
- They are willing to provide computing resources and dedicated personpower to support the project
- As for today legal matters are being sorted out
- We will need to drift towards transferability to other experiments

LHCb Brain

We are building team for the Proof of Concept (PoC)

Alexandre Boyer, Chris Burr, Ben Couturier, Ryunosuke O'Neil (LCB-CS)

Pawel Kopciewicz, Rosen Matev (sadly left CERN yesterday) (LCB-DP)

Micol Olocco (U. of Cincinnati)

Daniel Magdalinski (Nikhef)

Sergio and Carlomagno (CONARE)

+ possibly somebody else from DPA

During PoC we will need regular contact with analysts

17 research fellows in CERN LHCb group, as far as resources allow we will reach to entire LHCb

Physics priorities for PoC

We found 2 priorities to be the most appropriate:

1. In order to start a new Run3 data analysis, LHCb analysts wants to find all the HLT2 lines selecting a specific decay, and whether other analysis productions already select that information. They are then guided to create the Analysis Production code, validate and run it. This can be benchmarked with the help of the teams currently analyzing LHC Run3 data.

2. Make LHCb Brain able to understand an analysis note and related documents (that is reading text, analysis code, ntuples, presentations, …) at any point of its lifetime (started, in WG review, in review, …) to find discrepancies between the artefacts, inaccuracies, textual errors, and bugs as well as suggest improvements (This was exercised on 7 analyses so far and checked with their proponents, and 1 RTA paper in review with optimistic feedback.)

Frontier AI at LHCb?

Modern models like GPT 6 Astra or Fable 5.1 can perform very well on math problems.

So what's stopping them from searching for New Physics?

1. Leading AI companies generally do not fine-tune models on HEP data. HEP methods are a niche, and part of the knowledge is not public.
2. Analyses require hands-on interaction with internal CERN and LHCb services.
3. AI might challenge our established approaches and tools.
4. Not everything is documented.
5. We need CERN's agreements with AI providers on data privacy.

Frontier AI at LHCb?

Modern models like GPT 6 Astra or Fable 5.1 can do miracles on math problems.

So what's stopping them from searching for New Physics?

Knowledge Base + RAG + fine tuning

1. Leading AI companies generally do not fine-tune models on HEP data. HEP methods are a niche, and part of the knowledge is not public.
2. Analyses require hands-on interaction with internal CERN and LHCb services. **MCP servers**
3. AI might challenge our established approaches and tools.
4. Not everything is documented. **Find knowledge gaps and fill them**
5. Data privacy. **CERN contractual agreements with AI industry or MCP gateways**

Knowledge Base

The knowledge base (part of LHCb Brain) now consists of:

- LHCb CDS documents (papers, theses, internal notes, public notes, analysis notes, miscellaneous, reports)
- About 30 HTML domains (like Moore docs, LHCb main page, etc.)
- LHCb Indico since 2002 (includes institutes if accessible)
- CERN Twiki + Online TWiki
- Logbook
- Gitlab MRs + Issues
- Gitlab Code from LHCb software stack
- Gitlab Code from other LHCb repos (1014 repos) also scraped, however not kept in the knowledge base due to MCP server being more appropriate here. Also a lot of junk and noise code. And they take a tremendous amount of space.

Software projects in the LHCb Knowledge Base

Included documentation
lhcb/DBASE
lhcb/PARAM
gaudi/Gaudi
lhcb/Detector
lhcb/LHCb
lhcb/Lbcom
lhcb/Rec
lhcb/Allen
lhcb/Moore
lhcb/Alignment
lhcb/Online
lhcb/MooreOnline
lhcb/Panoptes
lhcb/DaVinci
lhcb/LHCbIntegrationTests
lhcb/Run2Support
lhcb/Boole
lhcb/Geant4
lhcb/DD4hepDDG4
lhcb/Gaussino
lhcb/Gauss

Web Crawler

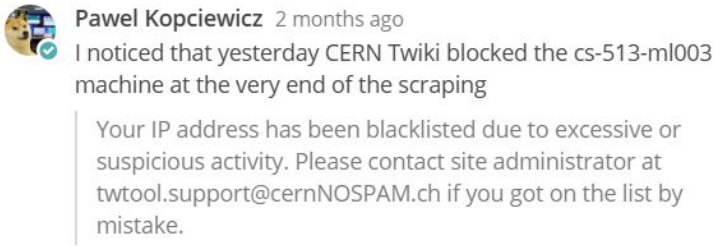
Web crawler walks and downloads every subpage of the domain

- We provide a list of domains, e.g. https://velo-u1-docs.docs.cern.ch/
- Scraped content is preprocessed to get rid of HTML syntax

The crawler includes TWiki (both online and CERN)

- TWiki format is very specific, we relied on AI to write preprocessor to get rid of all junk (this is a great example of difficult task AI is excellent at)

- Crawler generates load:



TWiki Pages

Web	Pages	Chunks
LHCbPhysics	2,895	38,750
LHCb	2,912	32,806
VELO	624	3,792
Online	571	3,065
LHCb/FAQ	48	1,080
L0	186	1,044
CALO	81	985
RICH	133	887
MUON	89	713
ST	50	505
OT	91	495
Operation	25	312
Computing	35	304
UT	41	293
Main	76	209
Testbeam	16	88
RP	16	68
SciFi	9	33
HLT	3	14

Scrapers

We have scrapers for CDS, Gitlab and Indico

- Each of them is a dedicated pipeline featuring different retrieval strategies (information retrieval from slides, PDFs, tables, formulas)
- They are quite complex software engineering problems
- Scraping is done using LHCb service account and Kerberos

Scrapers will need to be reviewed by relevant IT services

- Not hard to accidentally cause a DOS
- LHCb scrapers are cron jobs running in OKD (OpenShift) causing minimal load

What is RAG?

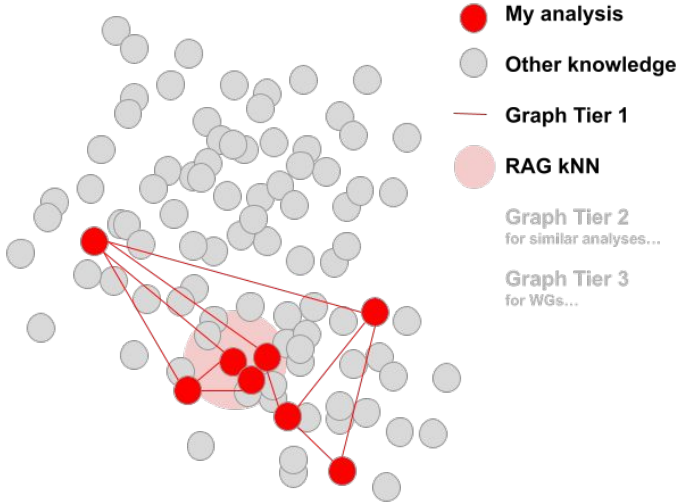
RAG stands for Retrieval Augmented Generation

The LLM answers using retrieved context from the knowledge base

Generalized concept:

1. We split documents into chunks
2. Convert each chunk into a vector (with the embedding model)
3. Store the vectors in a database
4. When user asks a question, we convert it to a vector and we find the chunks closest to that question
5. The LLM writes the answer from those chunks

Good retrieval means good answers!



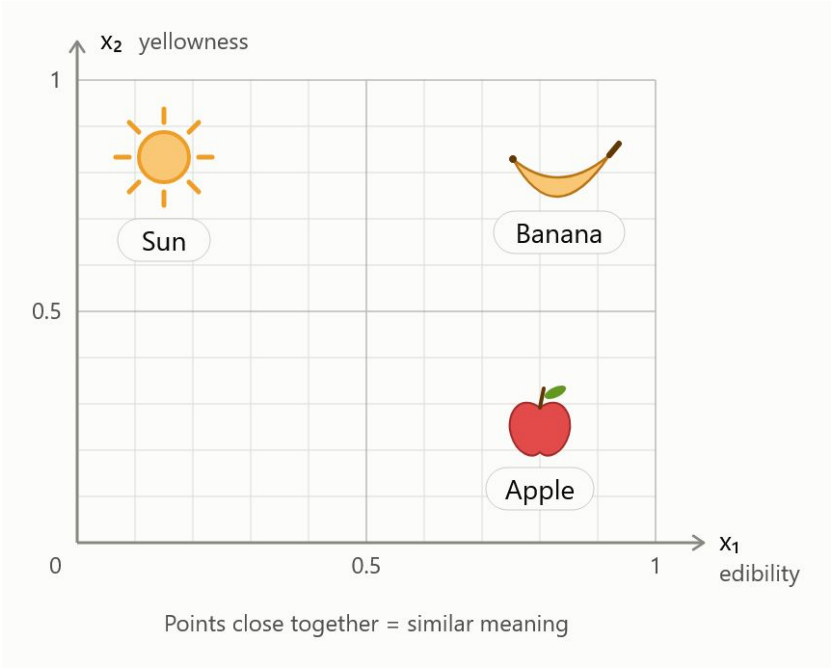
Knowledge in the vector space

Similar semantic meaning means nearby points

Each dimension captures some learned feature

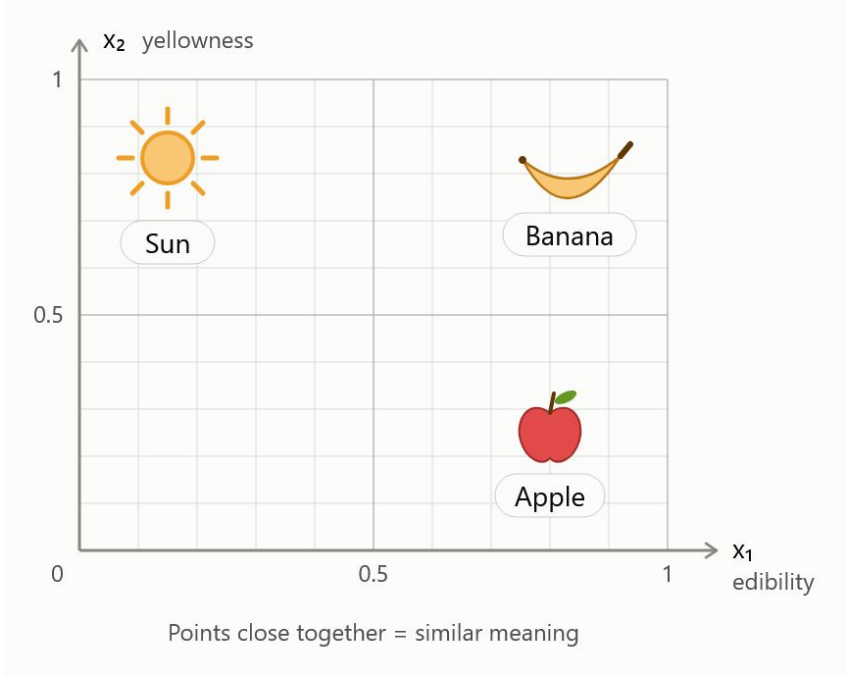
In real models there are hundreds or thousands of dimensions and they're not human-readable

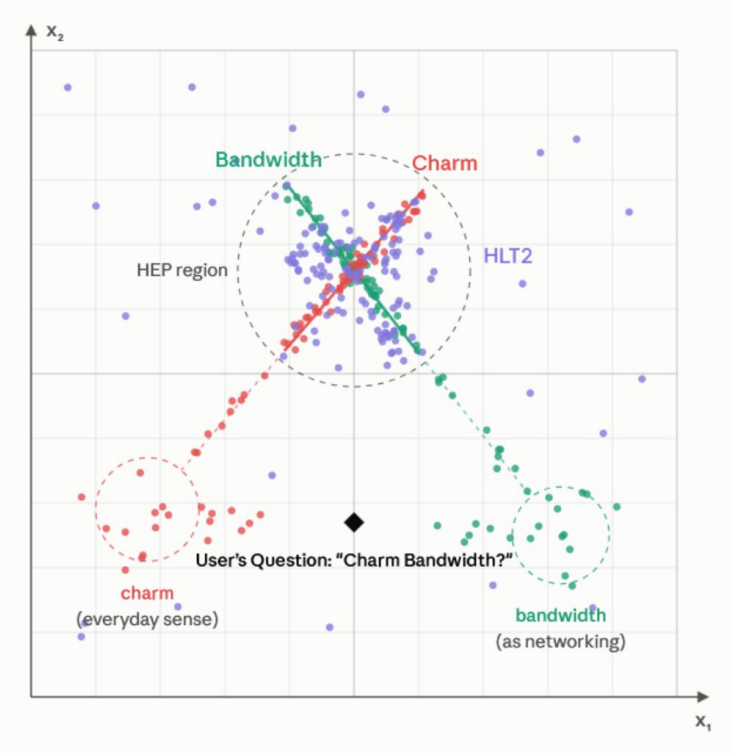
Vector space in general embedding model



Generic embedding model

Can we apply generic embedding model to HEP data? Yes, but…





PhD (and master) theses

Fact sheet about theses:

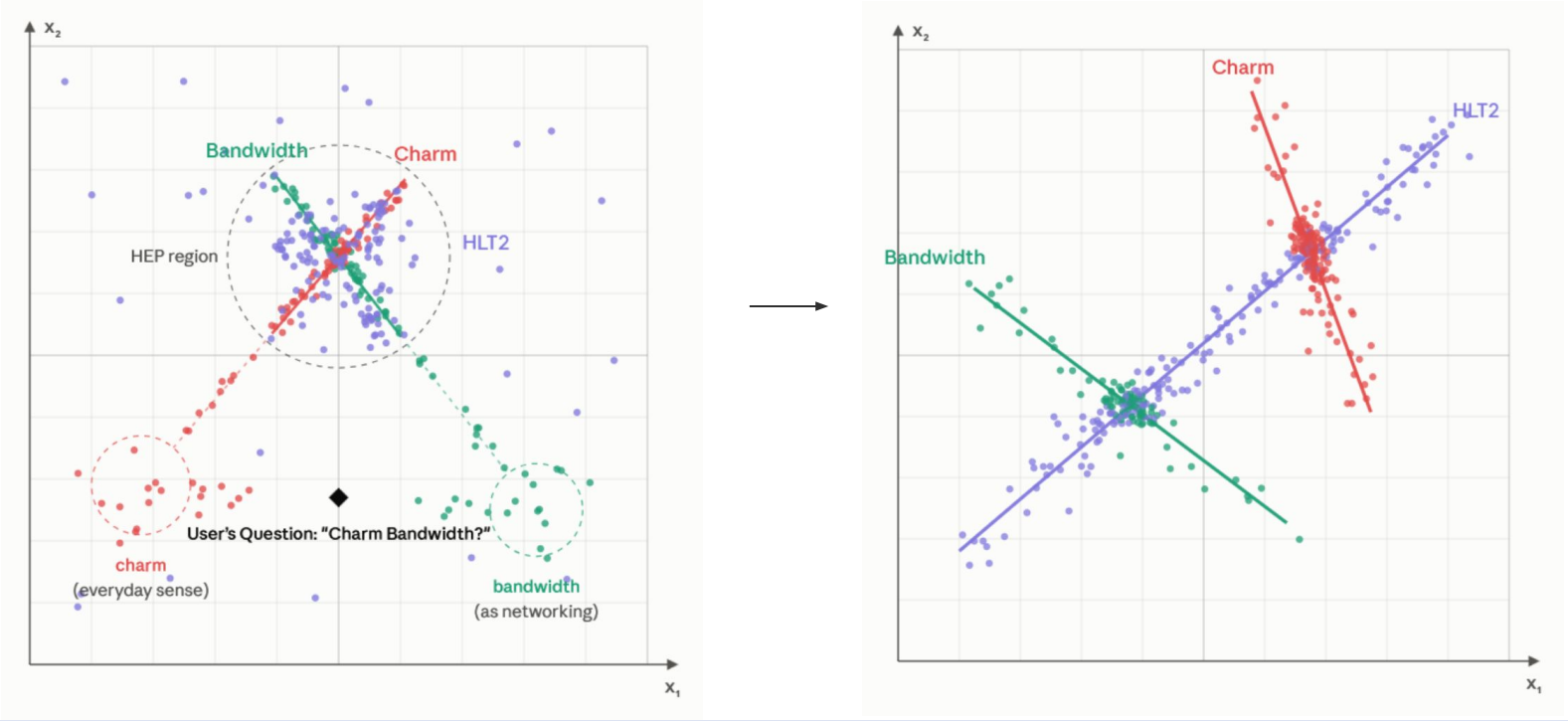
1. They are the biggest source of textual data in LHCb
2. They are almost never the best material to cite
3. Overlap between theses is massive, however every thesis is from unique perspective
4. They cover every area of LHCb, including internal jargon
5. They are public open
6. They are not peer-reviewed, may contain biased viewpoints, not accurate information, underdeveloped ideas, colorized phrasing, or even misleading information

All above makes them perfect for a training dataset!

Both for embedding model and for LLM

Vector space tuning

This is a visualisation of the perfect (probably not achievable) tuning



RAG in action

A question about a summary of a Charm meeting

Pawel Kopciewicz 11:14 AM

`/lbrain2 Hello, could you give me a brief summary of Charm meeting 2016, May 04? Please be concise and don't print any names`

Analysis worker 2 BOT 11:16 AM

On 4 May 2016 the Charm WG meeting covered two main analyses:

1. **ΔACP in Λc+ → p h h (pK+K−, pπ+π−)**
   - Status update on the ΔACP measurement using Run 1 data split by magnet polarity, with detailed selection and trigger requirements.
   - The analysis was reported to be progressing well with no major issues. The main open point was how to treat residual asymmetries after kinematic weighting in a robust way.
   - Discussion in the meeting focused on:
     - The need to match K+ vs K− and π− vs π+ spectra within each final state to control detection asymmetries.
     - Methods to estimate or bound residual nuisance asymmetries in the pK−π+ mode, possibly only needing an upper limit if effects are small.
     - The potential benefit of studying specific regions of phase space dominated by penguin processes rather than only phase-space–integrated observables.
     - Use of alternative variables (e.g. β asymmetry) versus mass under exchange hypotheses for vetoing reflections, and interest in Dalitz plots with pπ+ axes.
   - An analysis note was in preparation, with the goal of presenting results at the Charm 2016 conference.
2. **CPV searches in D0 → 4π with the energy test**
   - Follow-up on an earlier presentation, concentrating on tuning the metric parameter σ and validating the method.
   - A large D0 → K−π+π−π+ reference sample was split into subsamples; their p-values were found to be uniformly distributed and insensitive to production/detection asymmetries, confirming the robustness of the test.
   - σ was scanned for various CP-asymmetric scenarios; $\sigma \approx 0.5$ (GeV/c^2)^2 was chosen as default, with $\sigma = 0.3$ and 0.7 to be included in final results.
   - Backgrounds were studied using sidebands several σ from the peak; combinatorial and random soft-pion backgrounds were being incorporated stepwise.
   - The analysis is documented on a dedicated twiki and internal note; future work includes adding parity-sensitive variables and more detailed background studies.
   - In the discussion, it was suggested to repeat the reference-sample insensitivity tests for all σ configurations and statistics planned for the final result, and there was interest in starting WG review even before all parity-variable studies are complete, assuming their impact is small.

Administrative news included a reminder to apply for available conference talks (e.g. BEACH) and the announcement of the next WG meeting on 11 May 2016 in room 2-R-030.

Serena Maccolini 2:29 PM

> oh wow!

Mattermost bots

Slash Commands

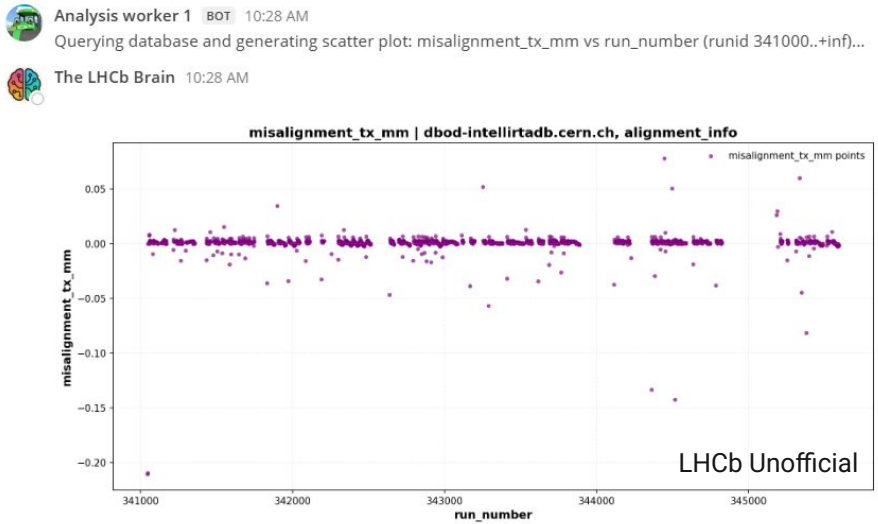
- Good for simple replies
- HTTP POST directly to the adapter in OKD
- They cannot post files, but a service account can be called to do it instead (see the picture)

Websocket Listener

- /api/v4/websocket connector listens to all events the service account has access to
- Bot token authentication enabled

However it is not possible to transfer SSO cookie through Mattermost endpoints

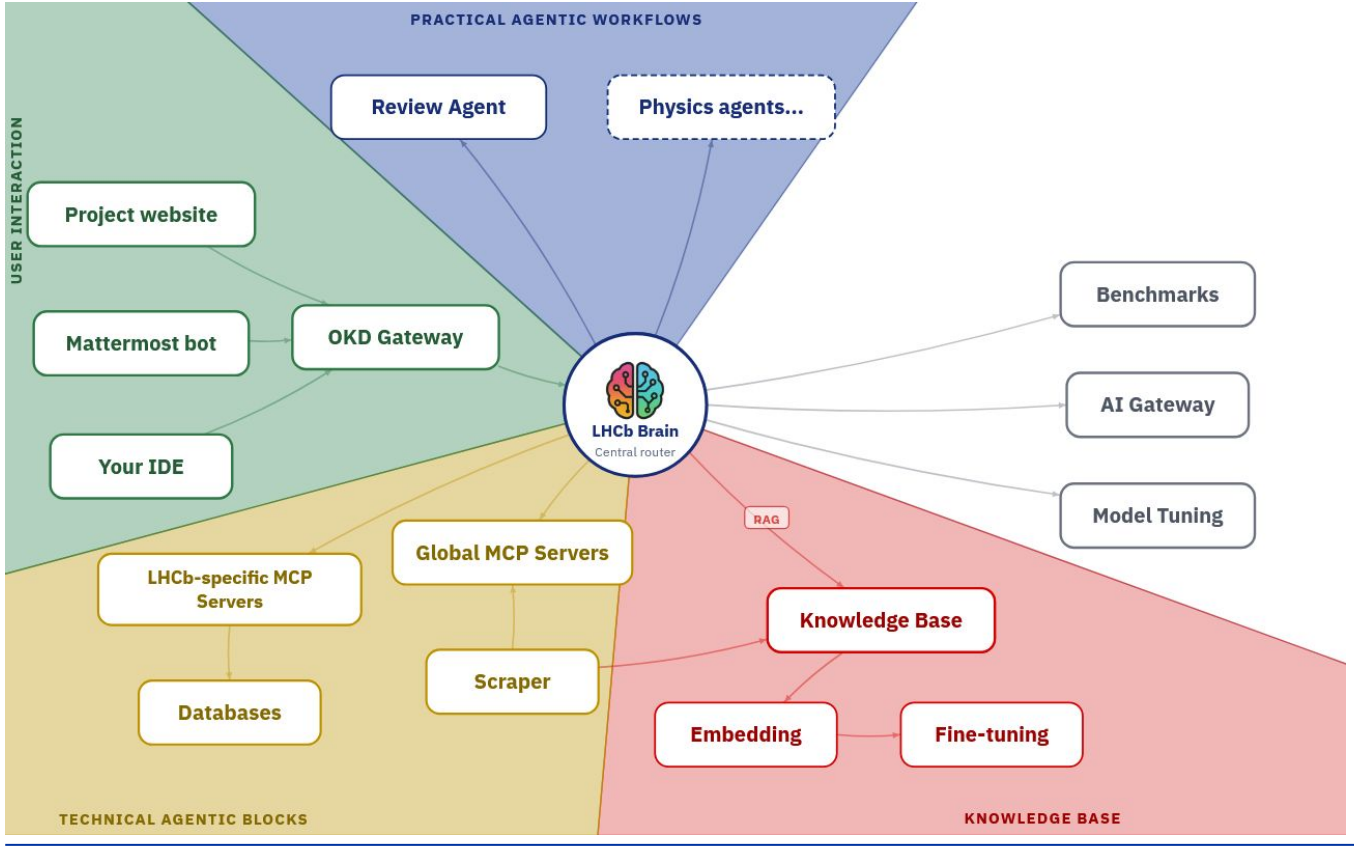
- For this we need our own chat application



The chain of routes underneath this Mattermost query:

Slash command -> OKD -> The Brain -> RAG -> The Brain -> MCP to PostgreSQL -> The Brain -> Plot Maker -> POST via Service Account

LHCb Brain overview



Chat Application

Our own application for agentic chats

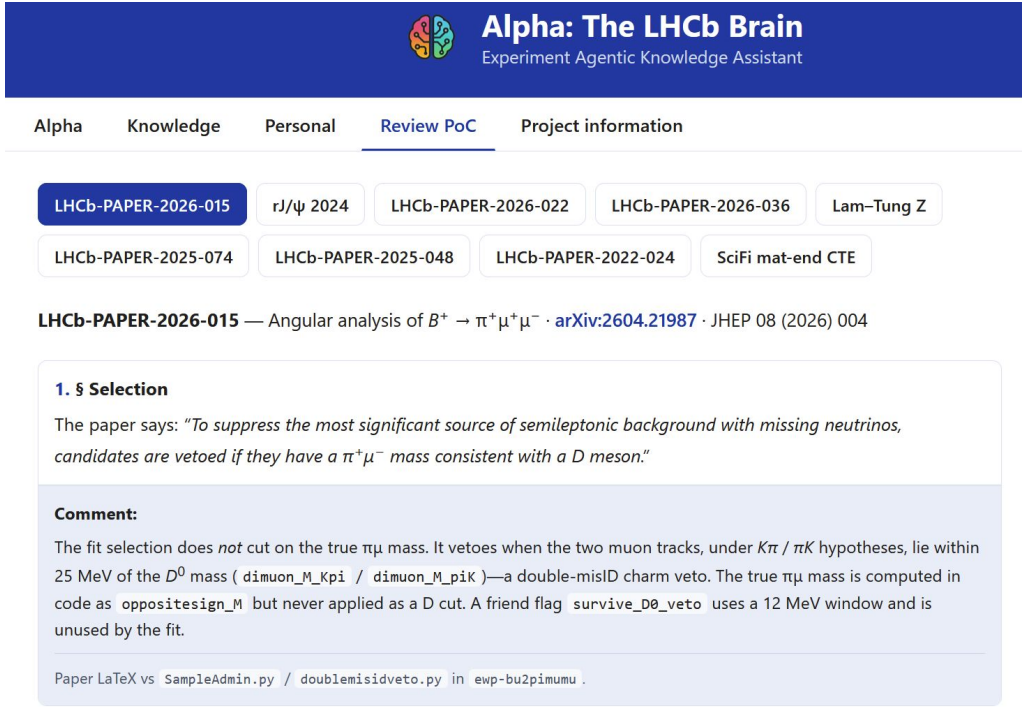
- Today it features chat tabs with selected KBs for our alpha testers
- The main advantage is that we can customize it for LHCb needs

What happens in the picture:

- We have a monitoring database with O(100k) entries, each chunk contains data locations, plot titles, labels, …
- The Brain calls RAG to identify possible candidates
- In this example, after RAG it was already clear what the user wants, so it queried numerical data from the MCP server (with a gateway to the experiment's online network) and called Plot Maker in the same query

Deep analysis review

This is alpha testers stage, not public to collaboration yet



One of panels of AI/ML discussions this Autumn (PC) will be about AI in paper review

Deployment, Services and Routing

OKD (OpenShift Kubernetes Distribution)

Our services:

- Core chatbot, router, RAG, …

  Due to resource quotas it runs locally with exposed API endpoints to communicate with services in OKD

- Cookie Refresher

  auth-get-sso-cookie-supreme running as a cron job; authenticates our service account via Kerberos

- SSO via Helm chart

  Pre-packaged Kubernetes configuration for deploying SSO

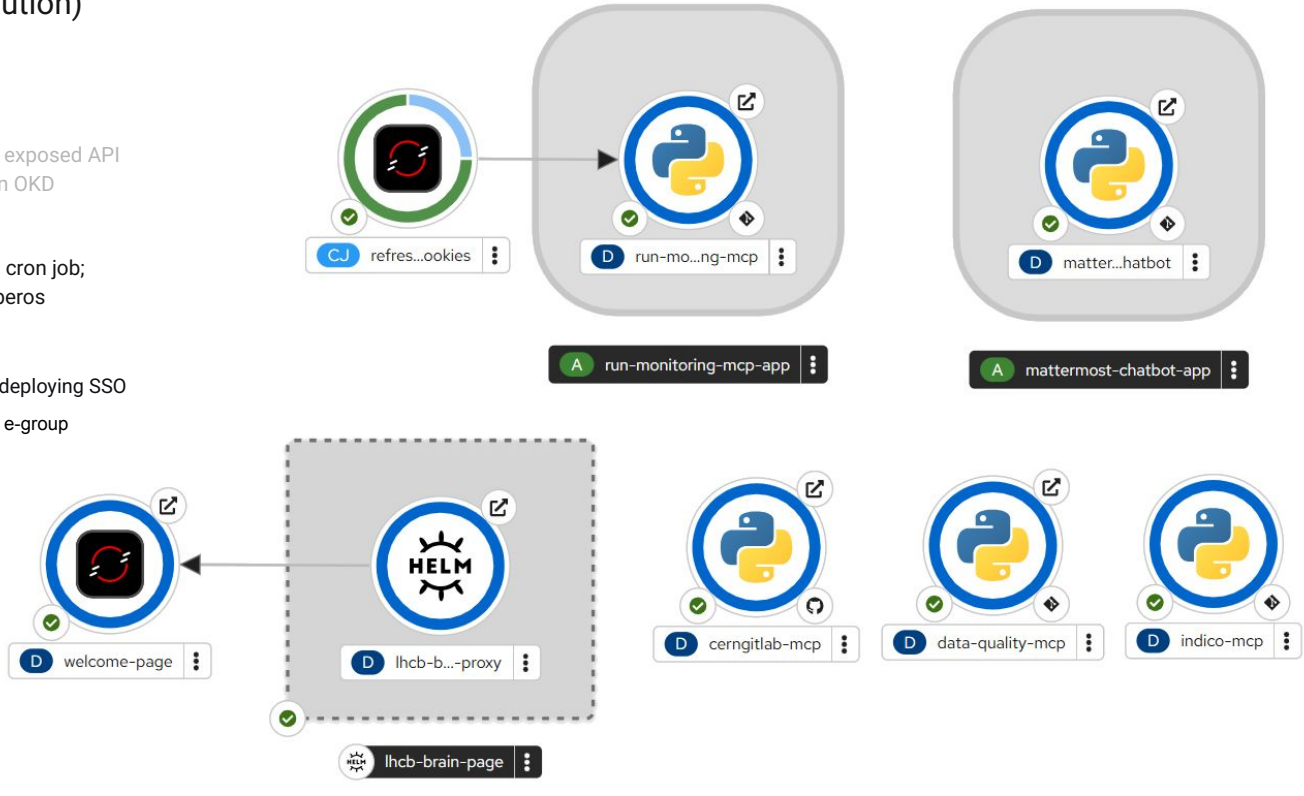
  Authorization depends on subpage (against an e-group membership)

- Multiple MCP servers

  They could become own projects

- Welcome page

- Mattermost Adapters



MCP servers

Every LHCb service with actions should have an MCP server

Read-only MCPs are 80% of success

This is because we can assume that the little time when a user has to follow as instructed can reduce chaos and be a natural point for reflection

(there will be exceptions)

Where we are currently staying

- Indico and Gitlab from central services
- Online (DQ, RunDB and OpenSearch) through IntelliRTA
- Developing Analysis Productions and related
- Developing OracleDB (archived WinCC datapoints)

MCP servers can be more specific - e.g. we could have one to review CI test

AI coding tools

CERN has conducted pilots of a couple of AI tools, among others:

- Gitlab Duo in 2024 (now available to all with limited resources)
- Cursor in 2025
- Claude ongoing, until Dec 2026

As for today, for general coding, in particular across the LHCb stack, my personal recommendation is a private Cursor subscription

Summary

- Lots of material I presented is an ongoing work
- Future is optimistic but uncertain
- If you are interested in AI topics you can follow the Agentic AI in LHCb Mattermost channel
- Please reach out if you have ideas (or doubts)
- The scope for AI in LHCb for 2027 is still not defined

Thank you for your attention

home.cern