NPTwins 2026 · ECT* Trento

LLM systems in high energy physics

From documentation chatbots to agents — what is being built, what works, and where it meets digital twins.

Gabriel Facini
University College London

11–15 May 2026

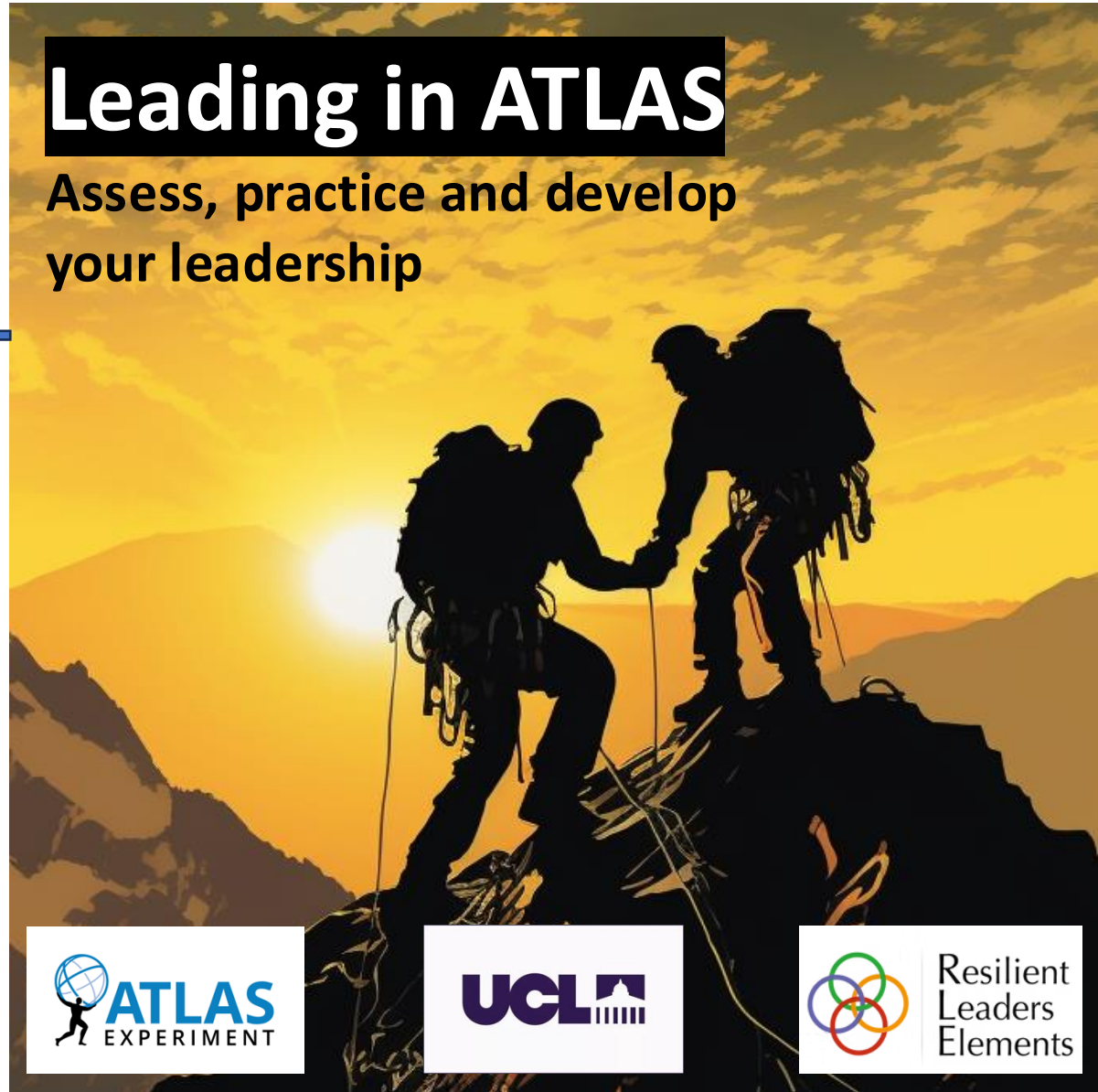
Please excuse my absence earlier in the week

To get more physics done, don't focus on physics

Leadership is a $350B business
CERN's budget is ~2B

A subset of our industry partners for PhD student projects

OliverWyman
PROTECHT AI
MissionZero
CERN openlab
lantum
Spotify
FATHOM
Climate Change Committee
UK Atomic Energy Authority
The Guardian
UCL Centre of Data for Intensive Science and Industry
Protect Group
UNDO PUTTING CARBON IN ITS PLACE
calda.ai
DiRAC
VORTEXA
Cabinet Office
NHS
nesta
SPACEFLUX
National Infrastructure & Service Transformation Authority
RevEng.AI
UKRI Science and Technology Facilities Council
Hartree Centre
Renbee



HEP experiments are tremendous



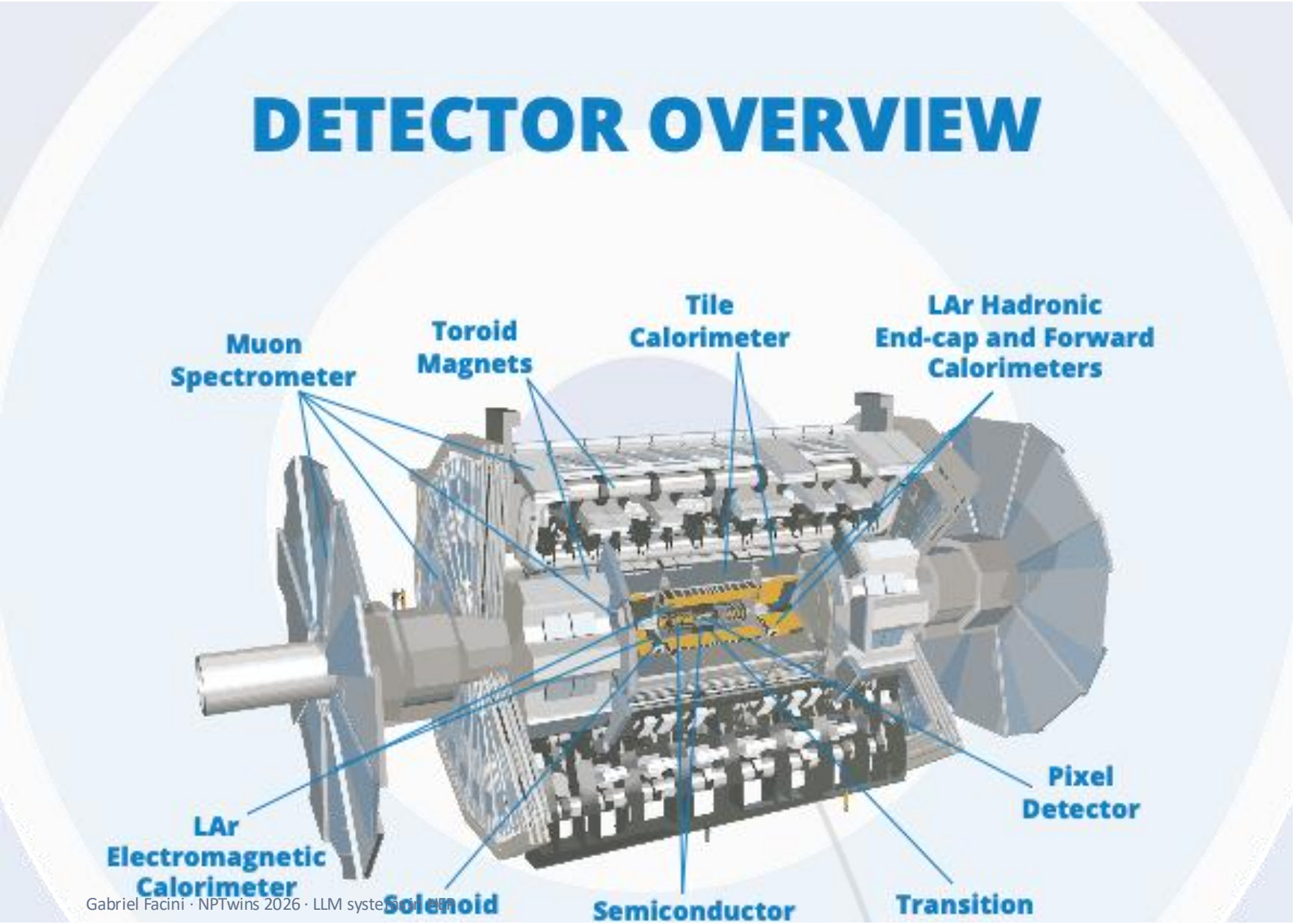
~6000 physicists, engineers, techs, students & admin

180+ institutions in **40+** countries

100+ nationalities

30+ years of existance

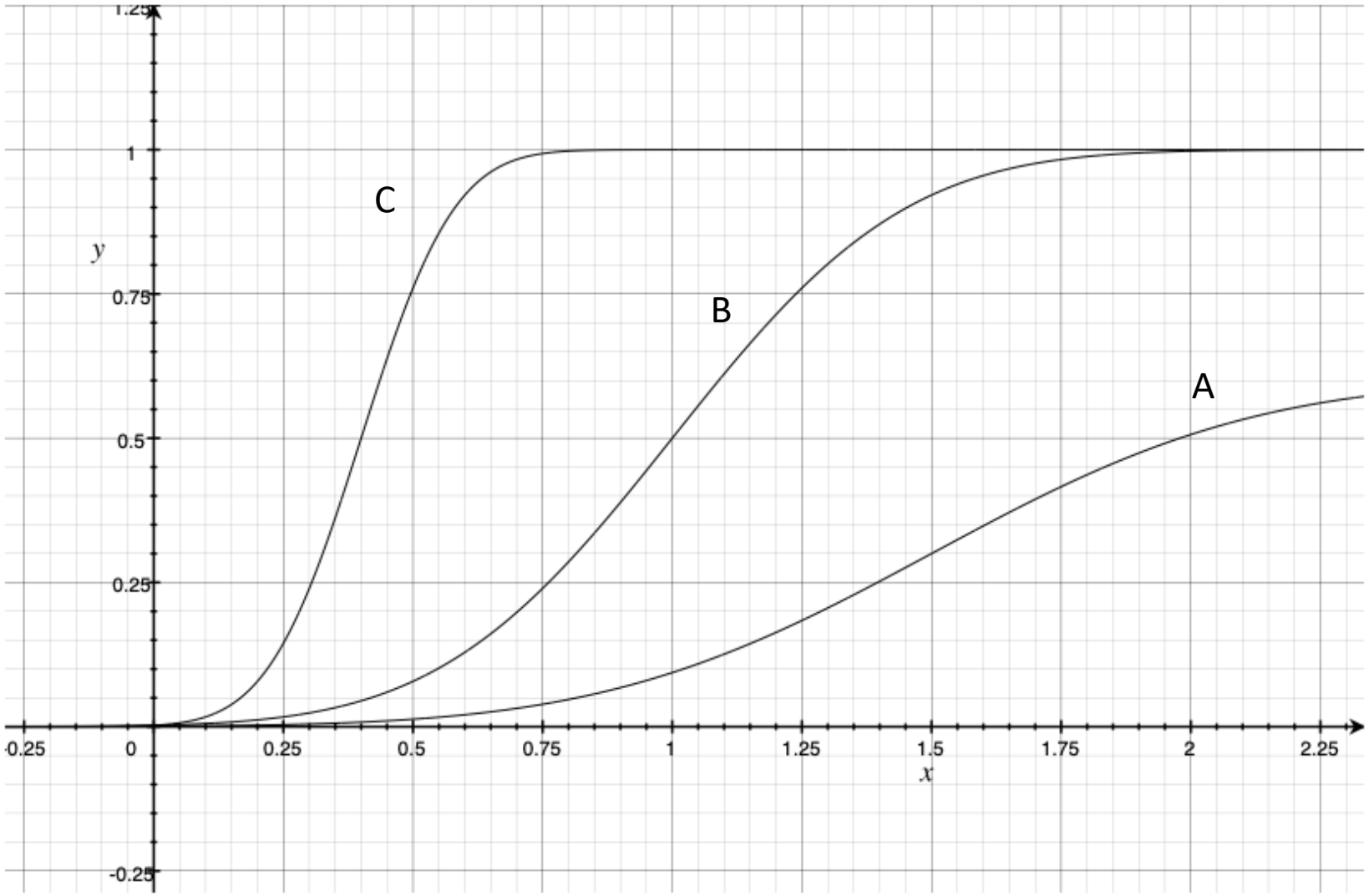
HEP experiments are tremendous



~100M channels
40 MHz of collisions
64 TB/s from ATLAS
1 kHz collected ~1.6 GB/s

Soon to be 5B channels!

How quickly until newcomers are proficient?



An experiment is more than its hardware

THE PHYSICAL / SIMULATED TWIN

Detectors, beam, electronics, reconstruction, MC.

- Petabyte-scale data
- Calibration & alignment
- Full simulation chains
- Reconstruction software

The classical scope of a facility-scale digital twin.

THE HUMAN / KNOWLEDGE LAYER

How a collaboration of 6,000 people knows what it knows.

- 35,000+ TWiki pages, 20 years of additions
- ~40,000 internal notes, 500,000 Indico events
- Mailing lists, chats, tickets, READMEs, slides
- Written in many voices, rarely curated

The gap that LLM systems are filling.

Embeddings: meaning as a direction in space

Words → vectors

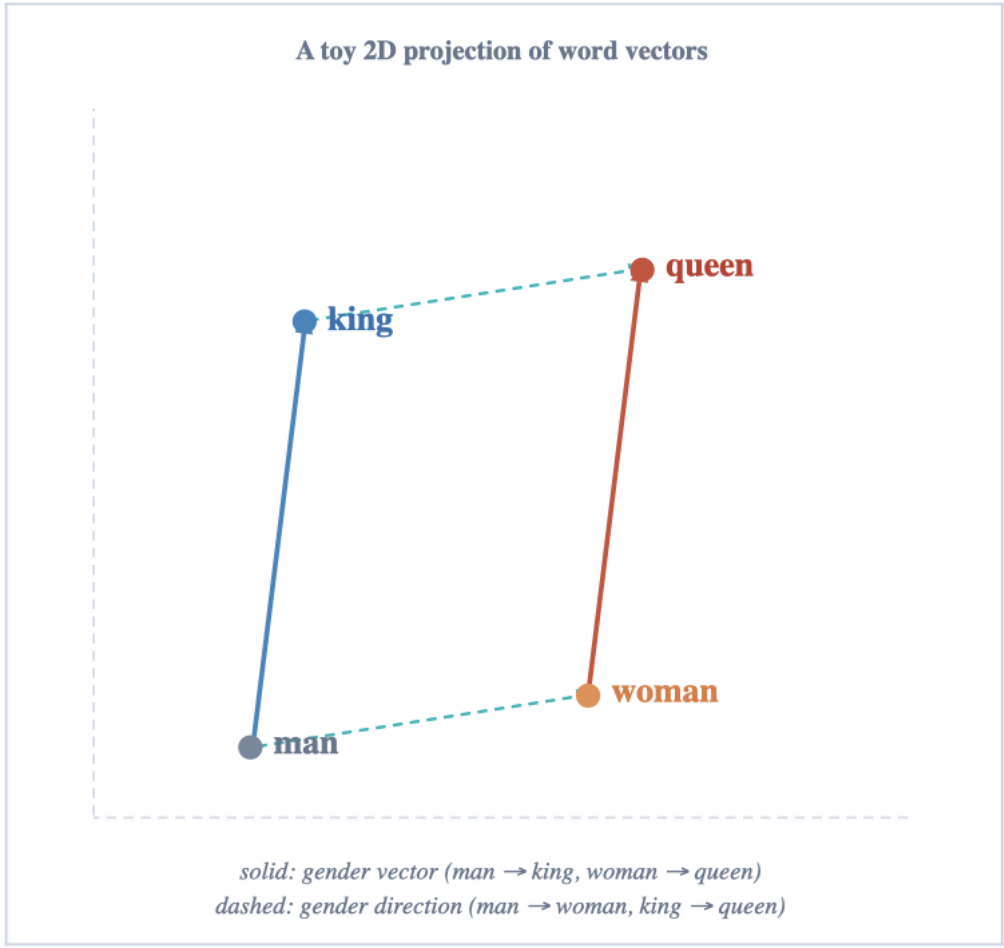
A neural encoder maps each word (or document chunk) to a fixed-length vector. Vectors that point in similar directions encode similar meanings.

king − man + woman ≈ queen

Vector arithmetic captures real semantic structure (Mikolov et al. 2013).

Why this matters for HEP

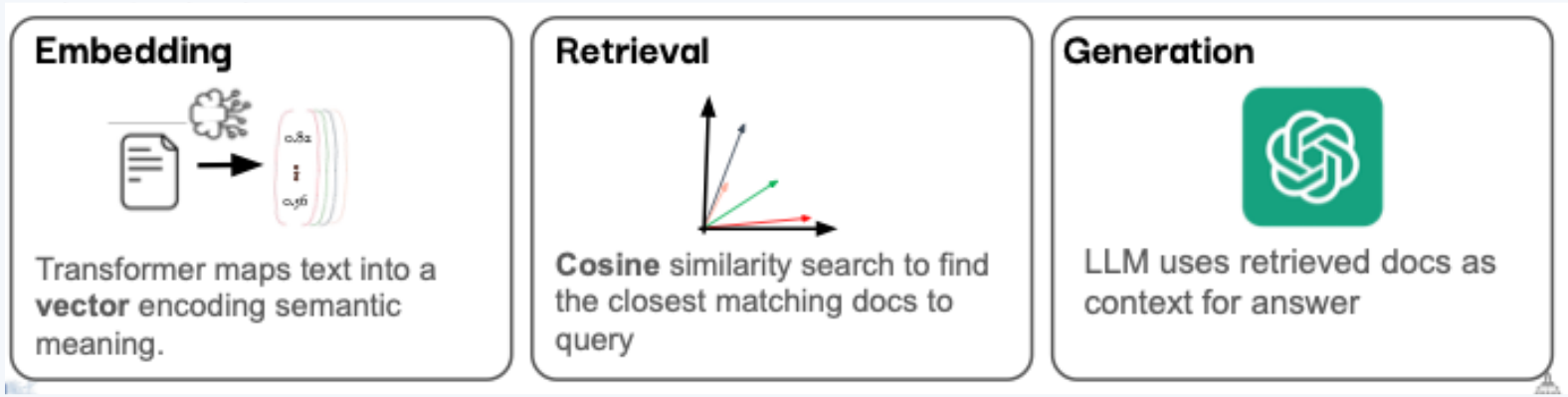
A user asking about "b-tagging" can match a document that only ever wrote "flavour tagging" — keyword search cannot do that.



chATLAS: Searching documents via embeddings

An example of Retrieval Augmented Generation (RAG)

Retrieval ~ parallelised search across all document forms and feed retrieved docs to LLM ffor answers grounded in internal material (outside the LLM training dataset).



$$\mathrm{sim}_{\cos}(\mathbf{q}, \mathbf{d}_i) = \frac{\mathbf{q} \cdot \mathbf{d}_i}{\|\mathbf{q}\| \, \|\mathbf{d}_i\|},$$

Why this matters for HEP

(Human) Resources are limited so we need to remove unnecessary barriers to entry.

*Juerg Beringer[1], *Joe Egan*[2], Ben Elliot[2], **Gabriel Facini[2], Daniel Murnane[1,3], Ki Ryeong Park[4], Alex Sopio[5]**, Samuel Van Stroud[2], Jeremy Couthures[6], Daniele Dal Santo[7], Joe George[2], Runze Li[8], Zhaohui Li[2], Cary Randazzo[9]

(1) BERKELEY LAB (2) UCL (3) UNIVERSITY OF COPENHAGEN (4) COLUMBIA UNIVERSITY (5) THE UNIVERSITY OF EDINBURGH (6) UNIVERSITÄT BERN (7) (8) Yale (9)

LLM systems are spreading across HEP

KNOWLEDGE ASSISTANTS

RAG over collaboration documentation

AccGPT
CERN accelerators
chatbot over CERN docs

chATLAS
ATLAS
hybrid retrieval, ~700 q / month

A2rchi
CMS-oriented
context-aware research support

DUNE-GPT
DUNE
knowledge retrieval & interaction

ANALYSIS AGENTS

Tool-using models that drive workflows

Dr.Sai (Xiwu)
BESIII
agent + domain LM for analysis

MARVEL
LIGO
multi-agent over unstructured data

Just Furnish Context
MIT / HEP
LLM agents for analysis tasks

MadAgents
MC generation
agentic event generation

OPERATIONS & CODE

Logbooks, control rooms, code intelligence

Logbook LLMs
Accelerator facilities
digest e-logs into insights

Code-aware RAG
ATLAS LXR, codebases
Future: code corpus integration

Copilot-style
Many experiments
code completion in IDEs

Routing/triage
Shift crews
early-stage experiments

Common backbone: retrieval-augmented generation. Divergence: how much agency the model is given to act.

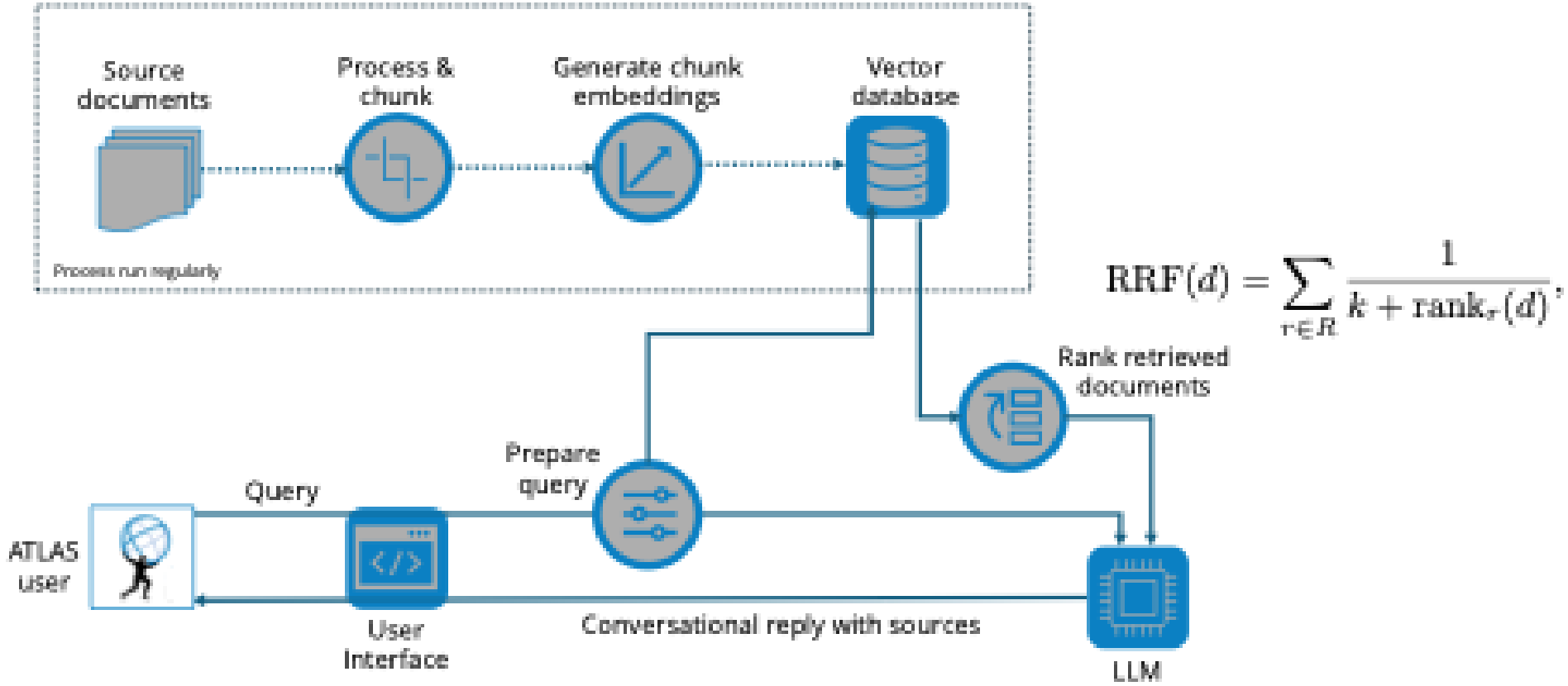
chATLAS pipeline

Don’t start from scratch

Find our code built on LangChain here

Table 1: Summary of dataset segmentation and retrieval statistics for the major ATLAS documentation sources. Parent and child chunks refer to the hierarchical segmentation used. N_{dense} and N_{sparse} refer to the number of sources returned from the dense and sparse retrieval, respectively.

Source	Parent Chunks	Child Chunks	N_{dense}	N_{sparse}
TWiki	148,707	661,153	10	5
MkDocs	45,838	201,681	10	5
CDS	224,965	1,029,826	10	5
ATLAS Talk	29,294	233,162	10	5
Indico	6,491	32,969	10	5

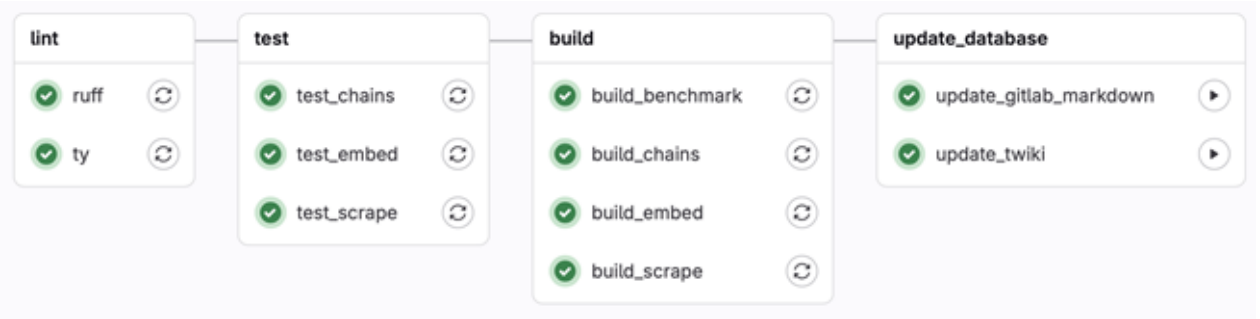


$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)},$$

Embedding: Storage and Automation

Storage

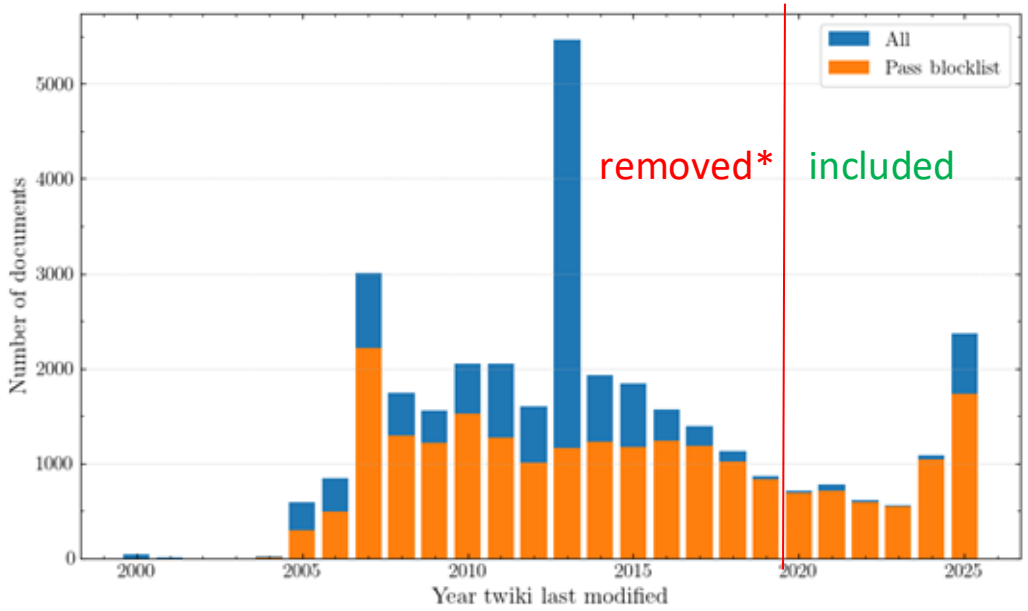
Embeddings hosted on CERN's Database on Demand service (DBOD) (postgreSQL) Stable and scalable



Auto-update pipeline

Gitlab CI/CD updates embeddings weekly

CDS, Indico, ATLAS Talk: jobs in progress



GraphRAG

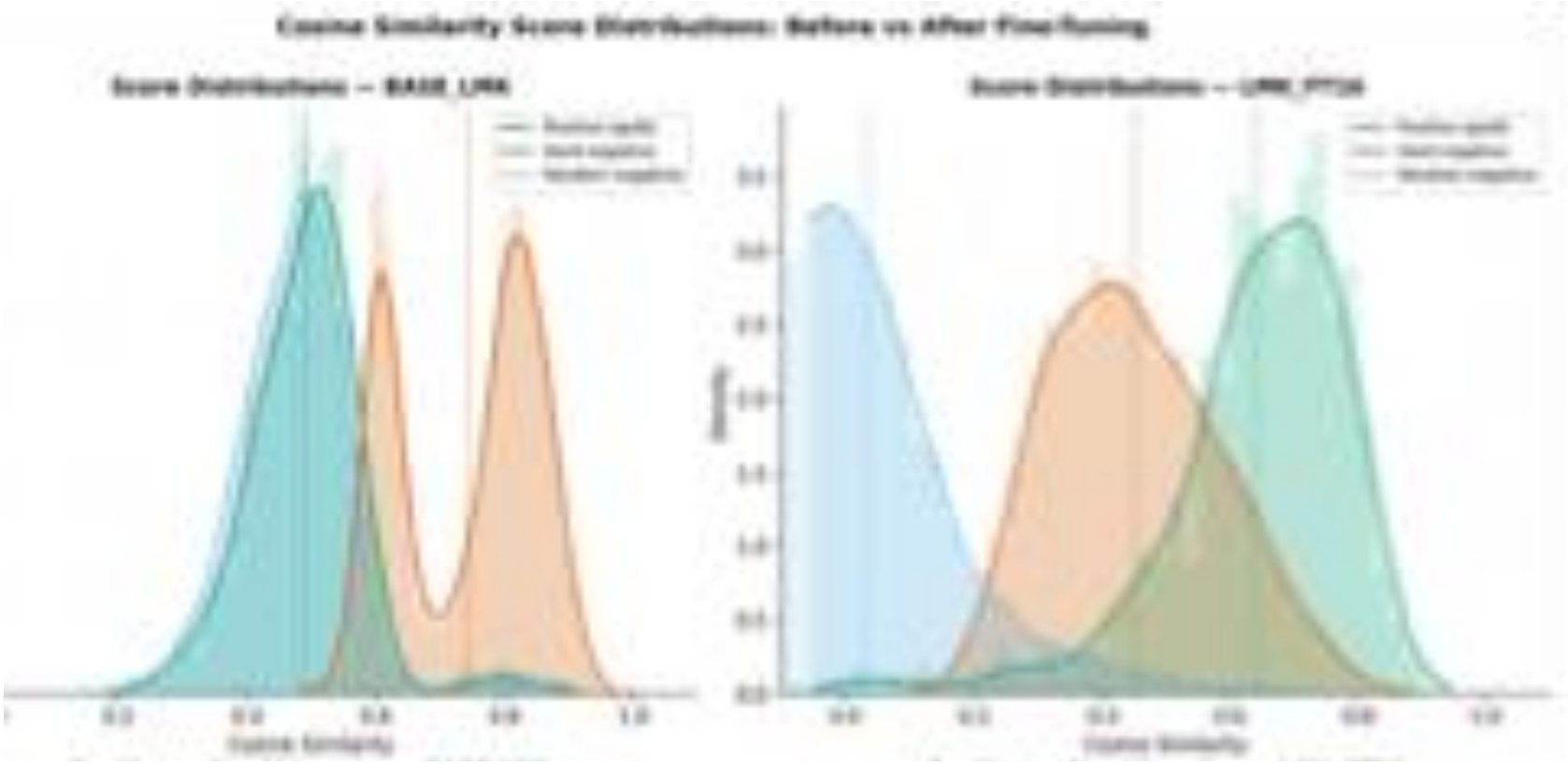
GraphRAG's are slow so depend on keywords to limit scope

ightRAG might be a useful compromise

Fine-Tuning Embedding Model

English words → ATLAS words

Find tuning sentence-transformers/all-mpnet-base-v2 on Question-and-Answer pairs helps distinguish semantic similarity in the ATLAS corpus



Credit: Richard Sutton, UCL MSci

Fine-Tuning Embedding Model

English words → ATLAS words

Find tuning sentence-transformers/all-mpnet-base-v2 on Question-and-Answer pairs helps distinguish semantic similarity in the ATLAS corpus

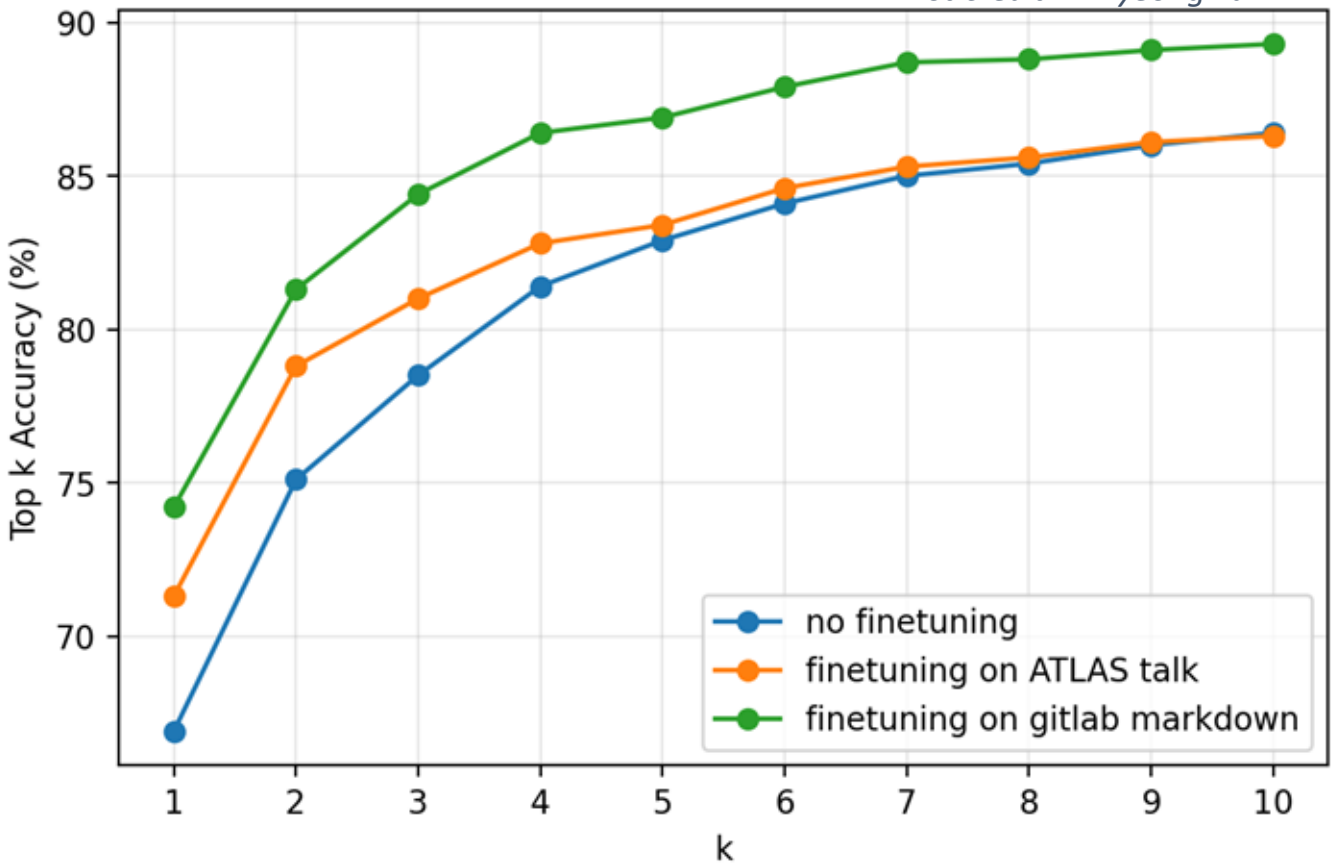
Q/A Pairs for tuning

ATLAS Talk is a Q/A Forum – Q/A in the wild

LLM-generated Q/A pairs otherwise (also for benchmarking)

Performance evaluated on Gitlab Markdown test set

Plot credit: Ki Ryeong Park



Fine-Tuning Embedding Model

English words → ATLAS words

Find tuning sentence-transformers/all-mpnet-base-v2 on Question-and-Answer pairs helps distinguish semantic similarity in the ATLAS corpus

Q/A Pairs for tuning

ATLAS Talk is a Q/A Forum – Q/A in the wild

LLM-generated Q/A pairs otherwise (also for benchmarking)

Q/A Pairs source matters

The more examples, the better, as general for AI/ML

Performance evaluated on Gitlab Markdown test set

Plot credit: Ki Ryeong Park

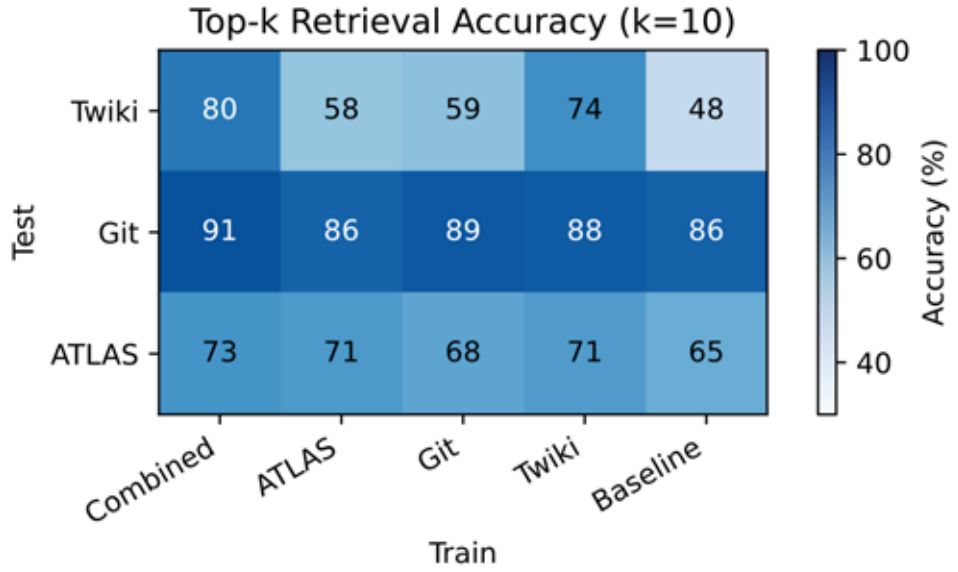


Figure 4: Top-k retrieval accuracy ($k = 10$) for cross-source embedding model evaluation. Rows indicate the test source, while columns indicate the source used for fine-tuning. "Combined" corresponds to training on all available QA pairs across sources, while "Baseline" denotes the untuned model.

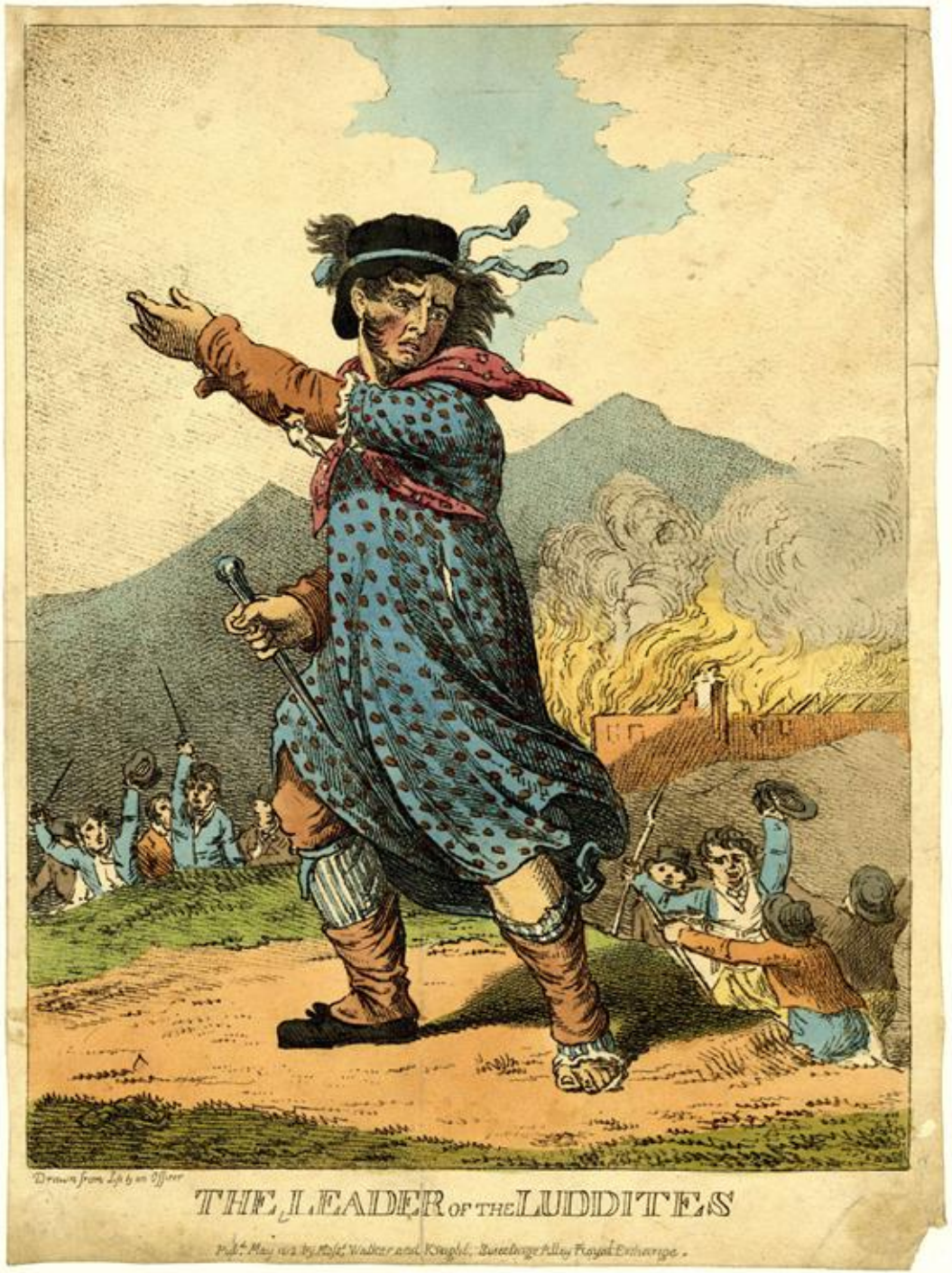
Who is the spokesperson?

Building trust in the system is a key element

Some people would rather see LLMs NOT be a useful tool – need avenues to build trust

Some people just want to be able to verify answers – provide sources/grounding





Access friction is solved. Mapping friction is not.

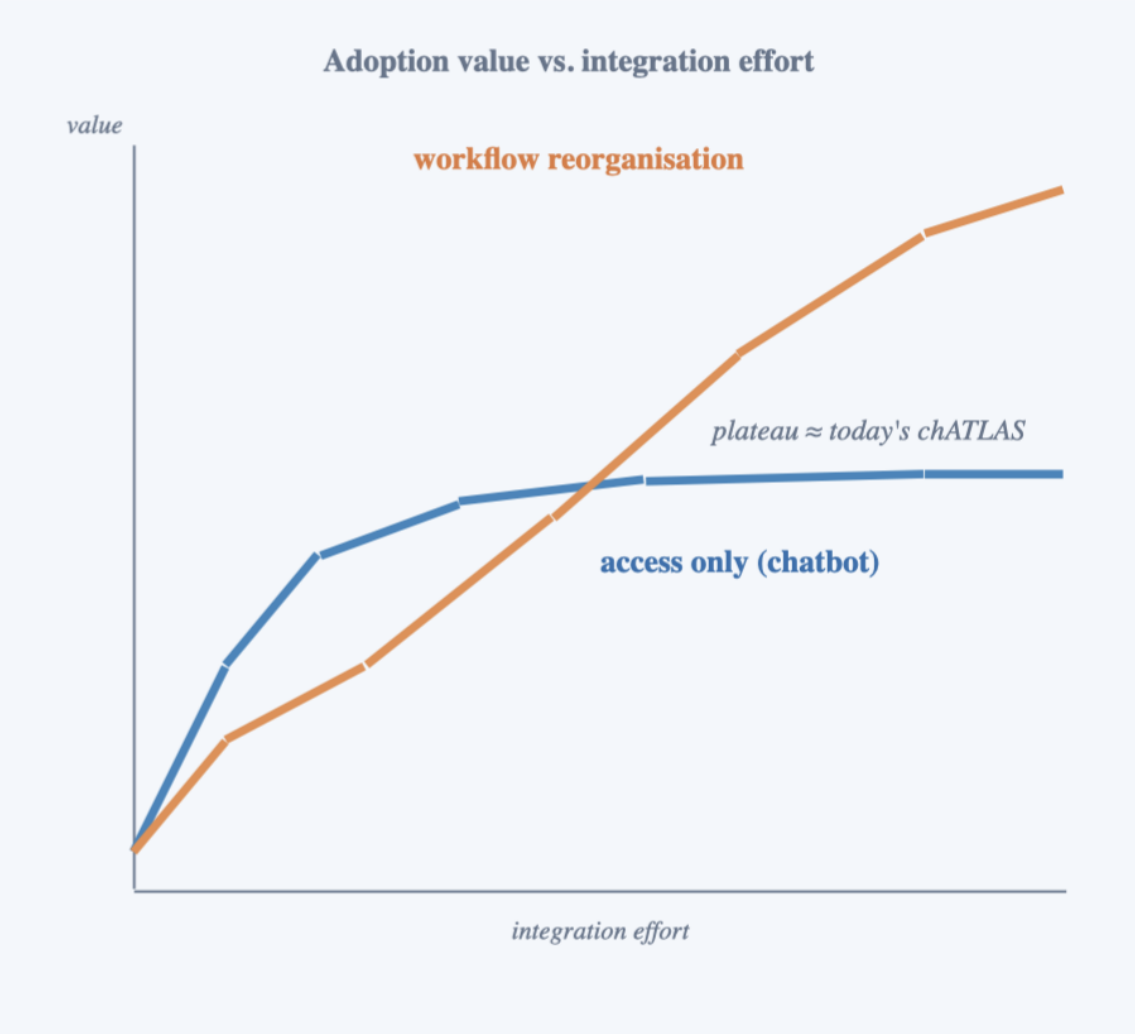
Two kinds of friction

Access friction

Can I find the right document at all? Solved by RAG: a chATLAS user gets a grounded answer in seconds where TWiki search would have failed.

Mapping friction

Do users restructure their work around the tool? The "mapping problem" (Kim et al.): in large collaborations, gains plateau quickly unless workflow chains are reorganised — not just individual steps automated.



Going agentic

As others have discussed in this conference…

Agents are becoming very useful especially on tasks like

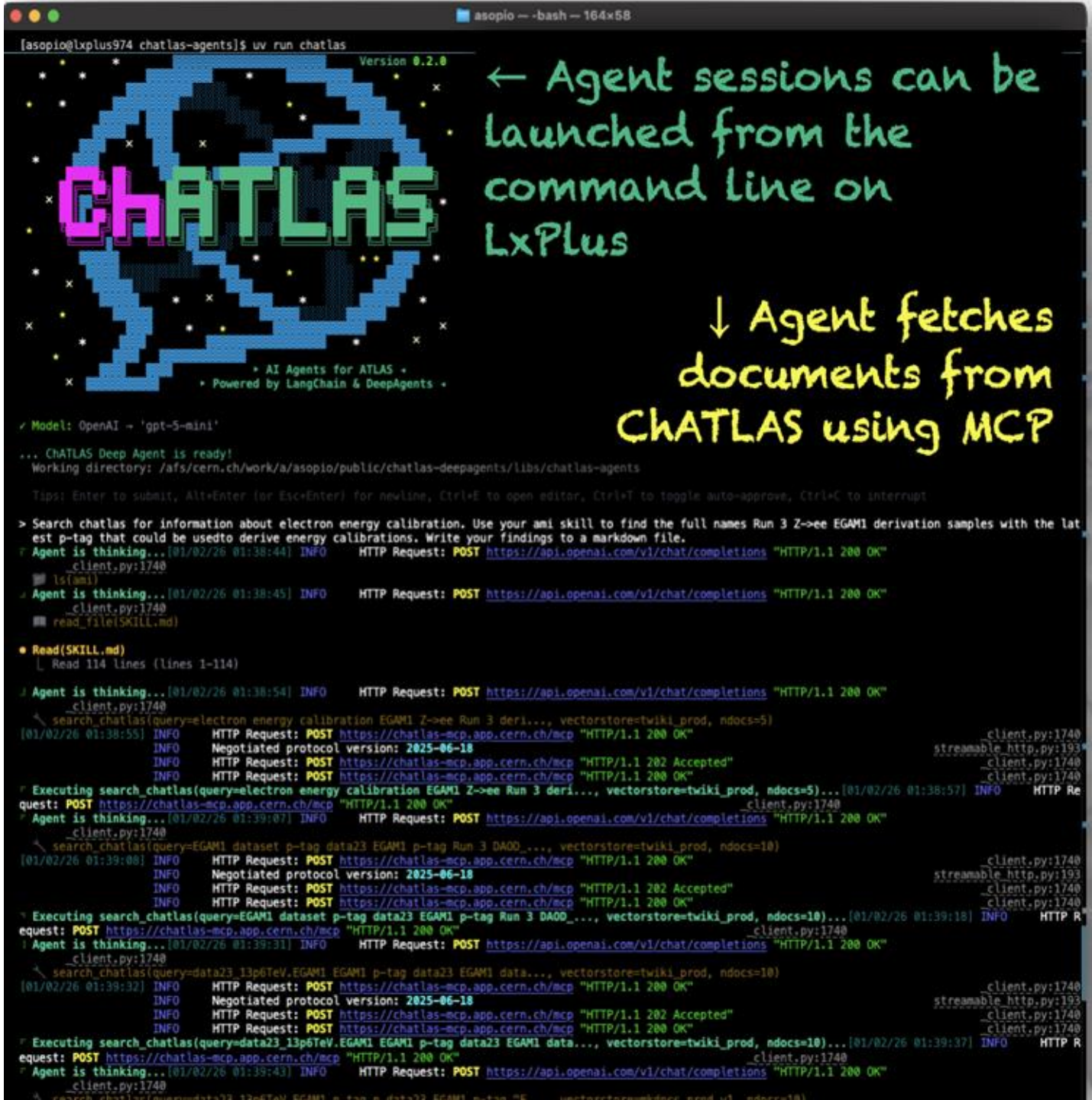
- Boiler-plate code
- Changes to verifiable code
- Debugging huge code bases

Building ATLAS specific pipelines

Automate, at least partially, some repetitive tasks.

Target: something agents will not do out of the box in 6 months

Usually hinges on us improving documentation



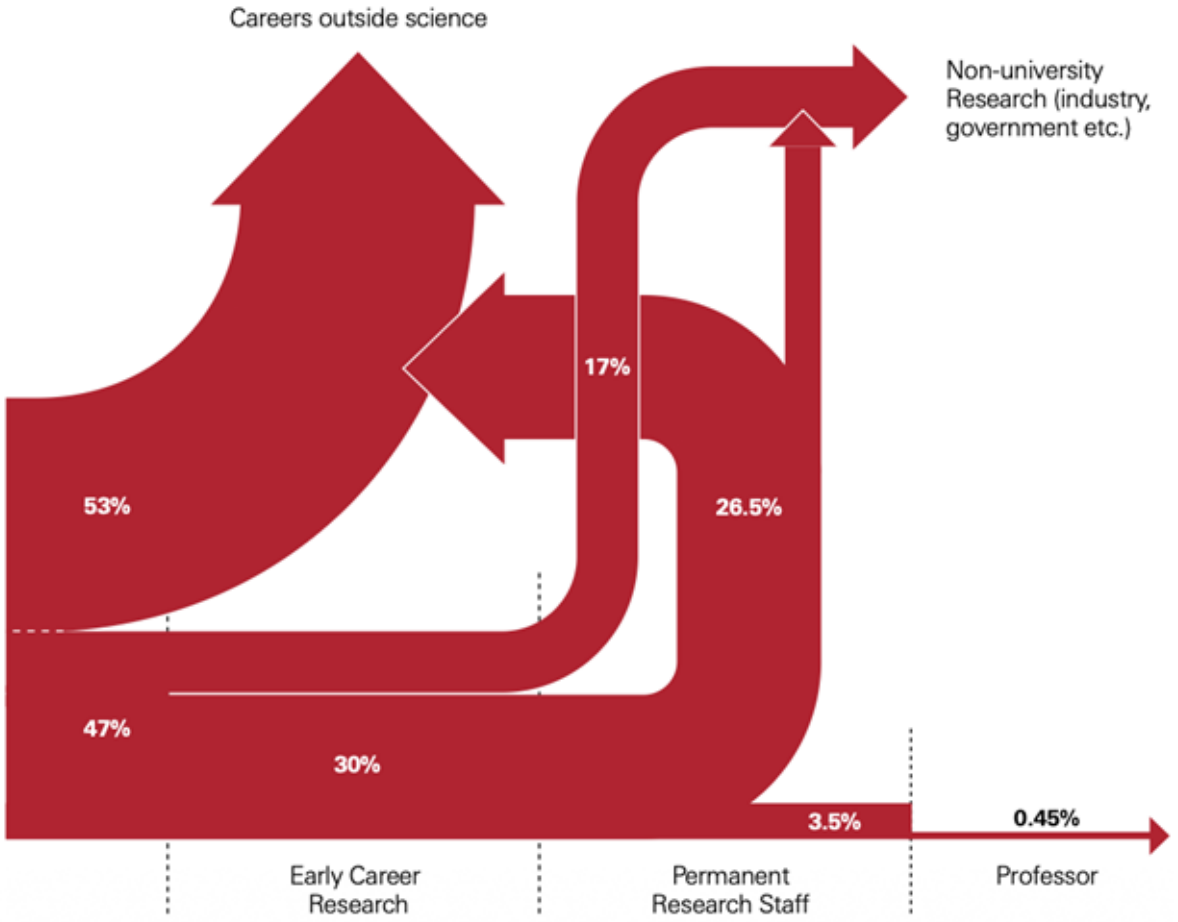
We love the science, but we are funded for training!

The Scientific Century
securing our future prosperity

Richard Feyman Quote

Science is like sex: sometimes something useful comes out, but that is not the reason we are doing it

Figure 1.6 **Careers in and outside science**



Unique Skillsets

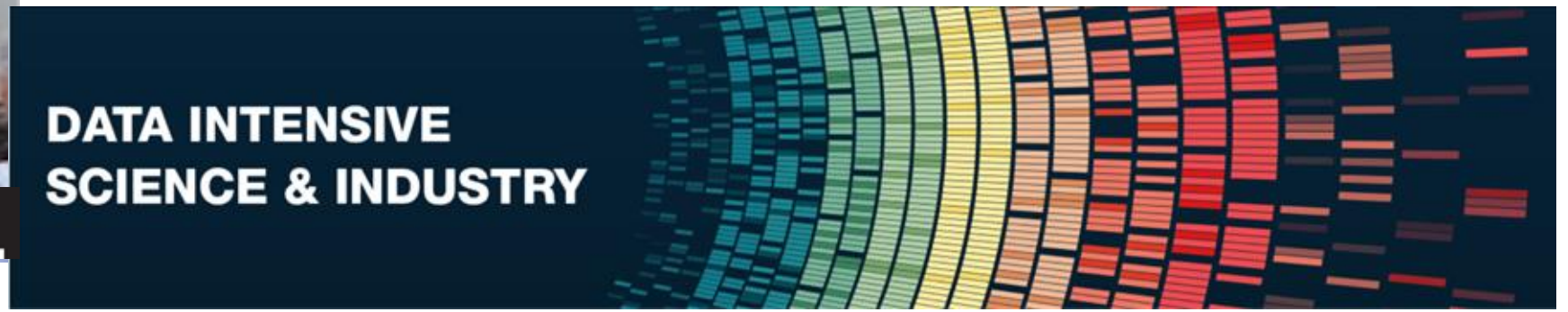
Joe Egan → chATLAS lead developer

LLM familiarity started with an industry project with The Guardian, project eventually this article.

Developed skills further making chATLAS production-ready.







Where LLM systems and digital twins meet

PHYSICAL / SIMULATED TWIN

Detector, beam, MC, reconstruction

The digital twin's classical scope — physics models bound to instrument data.

↓

KNOWLEDGE & WORKFLOW LAYER

Documentation, code, tickets, decisions

The institutional memory of the collaboration — fragmented and rarely indexed.

↓

NATURAL-LANGUAGE & AGENT LAYER

Retrieval, reasoning, tool use

LLM systems sit here. Through MCP they can read the lower layers and act on them.

THE OPPORTUNITY

A digital twin without a human-interface layer is sub-optimal.

- Query simulation outputs in natural language
- Trace a calibration decision to its document
- Drive workflows through MCP-exposed tools
- Lower the entry barrier to facility-scale DTs

LLM systems are not the digital twin — they are how humans and other agents reach into it.

Take-aways

1 RAG works in HEP

Domain-adapted embeddings, hybrid retrieval, and traceable sources are now in production.

DT communities can benefit in the same way!

2 Finetuning has large gains

Our language is specific, and the same information used to benchmark a RAG is exactly what can be used to fine tune it.

3 Agents are the next layer

MCP exposes documentation, code, and operational tools to LLM agents — turning chatbots into orchestration layers over the experiment.

4 Who are we training for what?

Opportunities exist to enhance the skillset and better prepare people for life after degrees (or contracts) end.