chATLAS

Towards an AI assistant for the ATLAS Collaboration

Joe Egan OBO The chATLAS Developers

UK Meeting on AI in HEP, Liverpool

17th September 2025

ATLAS
FUELED BY ML/AI

UCL CDT DIS

LLMs

- Already extremely prevalent in both research and industry
- Generative AI can automate tasks not thought possible 5 years ago
- Well known drawback: hallucination

A few HEP use cases:

- Coding
- Editing/Reviewing papers
- Efficient information retrieval → *focus of this talk*

Talk Plan

1. The problems we’re trying to solve
2. What we’ve done
3. Where we’re going

Two problems

Finding what you need

- **Documentation** is the primary medium for knowledge transfer in large collaborations
- Information you need is spread over myriad forms (wikis, papers, slides, markdown files…)
- Everybody has to search the docs

An efficient **information retrieval tool** can benefit the entire collaboration

The elephant in the room

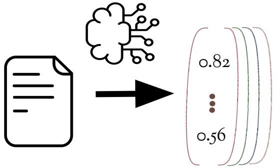
- LLMs are powerful generalist tools
- **Can struggle in specialised domains** like HEP, especially since they aren't trained on internal documentation
    - Fine-tuning a full scale LLM not feasible in the short term

A LLM allows you to find answers/information in a **conversational way**

Two problems: One solution

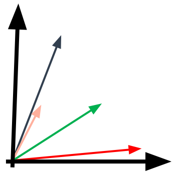
- Both can be addressed with Retrieval Augmented Generation (RAG)
- Retrieval ~ unified search engine across all document forms
  - A useful tool in itself
- Feed retrieved docs to LLM → better answers
- This is a Natural Language Processing task. Apply **industry experience** from my CDT group project

Embedding



Transformer maps text into a **vector** encoding semantic meaning.

Retrieval



Cosine similarity search to find the closest matching docs to query

Generation

LLM uses retrieved docs as context for answer

Example: chATLAS

What are the run 3 recommendations from the muon CP group for my ATLAS analysis?

28 results

Run-3 muon recommendations from the Muon CP (MCP) group

1. Muon identification – use the `MuonSelectionTool`

- The tool implements the official Run-3 working points (WPs):
- **Tight** – combined muons
- **Medium** – combined muons
- **Loose** – combined muons
- **High-pT** – combined muons (intended for very high-pT muons)
- The selection applies the standard η acceptance, quality cuts defined by the WP, and the required ID-track criteria (hits, impact-parameter cuts, etc.)

2. NSW as a precision layer

- For releases **≥R24.2.24** and with a p-tag **≥p5834**, the New Small Wheel (NSW) can be used as an additional precision layer in the identification step.
- In this configuration the end-cap **High-pT** WP is disabled (the NSW-based fit is already optimal for high-pT tracks).

Example: chATLAS

3. High-pT muon selection

- Analyses that contain a non-negligible fraction of muons with $p_t > 300$ GeV must use the **`HighPt`** WP (or obtain explicit approval from the MCP conveners).
- The `HighPt` WP is tuned for the larger multiple-scattering and alignment uncertainties at very high momentum.

4. Trigger scale factors

- Apply the Run-3 muon trigger scale factors (SFs) provided by the MCP group. The latest SFs are available on the MCP Twiki page *"Muon Trigger Physics Recommendations Run 3 (2022)"* (link in the documentation). These SFs should be used for both single-muon and multi-muon trigger efficiencies.

5. Performance references

- Muon reconstruction efficiency and momentum-resolution performance have been measured with the full Run-3 dataset. Public plots and the accompanying ATLAS-COM note (ATL-COM-PHYS-2022-827) can be consulted for validation of MC-to-data corrections.

6. Practical implementation

```
# Example configuration fragment (Athena/TopCPToolkit)
MuonSelection:
  tool: MuonSelectionTool
  workingPoint: Tight   # or Medium / Loose / HighPt
  useNSW: true          # only for releases >= R24.2.24 & p-tag >= p5834
  applyEtaCut: true
  applyTrackQuality: true
```

Sources Used by Model

MuonPerformancePublicPlots

atlasphys-top/docs/top :: Reco/Recommendations/run3_R25/muons

Some of the challenges

details in the backup

Sources	Embeddings	LLM	Benchmark
Getting the latest and greatest Variety of sources • Twiki • Gitlab Markdown • CDS • Indico • ATLAS Talk (Q&A forum) Putting together pipeline to *auto-update*	**Map HEP language to vectors** • Fine-tuned embedding model on ATLAS Q&A pairs **Stability and scalability** • Move to [CERN-hosted] postgres server for embedding storage	**Scalable LLM service** • Solution from CERN IT: Access to suite of low latency models via Gro**q*** [API] • Order of magnitude lower latency than OpenAI	**Making sure it actually works** • LLM-generated, human evaluated Q&A set • Used to evaluate design choices

**Groq*, not Grok*

Looking Ahead: Going “Agentic”

- Agentic = A language model that can call tools
- Two types of agency relevant for chATLAS

1. **Agency within RAG, examples:**
    - LM routes to appropriate source
    - Evaluation of answer before reply

2. **Agency to call other tools**
    - Document retrievers just one type of tool
    - Can add other tools (functions) a LM can call to complete a task
        - Dataset discovery, detector conditions etc.
    - Director LLM calling smaller, specialised models

SLMs?

Small Language Models are the Future of Agentic AI arxiv:2506.02153

Peter Belcak[1] **Greg Heinrich**[1] **Shizhe Diao**[1] **Yonggan Fu**[1] **Xin Dong**[1]
Saurav Muralidharan[1] **Yingyan Celine Lin**[1,2] **Pavlo Molchanov**[1]
[1]NVIDIA Research [2]Georgia Institute of Technology
agents@nvidia.com

- Currently in the Large Language Model paradigm
- Some advantages of SLMs:
  - Faster, much more economical
  - Realistic to fine-tune
  - Can build composite agentic systems from many specialised SLMs → “mixture of experts”
- Much more investment in **L**LM training infrastructure so far

Concluding Remarks, Discussion Points

Summary

- Primary goal of chATLAS: **To help people find information more quickly**
  - Can increase efficiency of whole collaboration
- We have a working tool
  - Lots of effort has gone into ensuring it is scalable and the codebase maintainable as the project grows
  - Still a good amount of low-hanging fruit

To Discuss

- RAG other HEP experiments
  - See examples here and here
  - Can benefit from shared tools, open source code
  - Limited by the quality of documentation you give it (garbage in, garbage out)
- Other tools
  - e.g. Coding, paper editing
  - Many more LLM calls, so **low latency** LLM service is **crucial**
  - *Could the UK provide this?*
- Training/finetuning a LM
  - Significant resource requirements, SLMs a way forward?

Backup

Where we’re going: Near term

Sources	Embeddings	LLM	Benchmark
• Scheduled scraping and database rebuild pipelines for all sources	• Fine-tuned embedding model for each source • Optimisation of postgres DB (nontrivial to balance latency and accuracy at scale)	• Real-time LLM as a judge answer evaluation, to refine search if needed	• Benchmark questions covering all sources • Adding real user questions to benchmark (“truth” document can be hard to find)

Example: OpenAI

What are the run 3 recommendations from the muon CP group for my ATLAS analysis?

Thought for 8s

Nice — here's a compact, actionable summary of the Muon Combined Performance (MCP / "muon CP") **Run-3 recommendations** you should apply to your ATLAS analysis, with links to the official guidance I found.

Short summary — what to do

1. **Use the latest muon reconstruction & ID WPs** (Loose/Medium/Tight) and the muon CP algorithm sequence recommended for Run-3. These were updated for Run-3 and the CP alg sequence was pushed into Athena (use the recommended alg chain in your reconstruction/config). gitlab.cern.ch +1
2. **Apply the MCP-derived scale factors and momentum calibration (including sagitta-bias corrections)** to data/MC. The MCP group provides calibration constants and sf's for efficiencies, momentum scale and resolution that you must apply to get correct yields and shapes. Indico +1
3. **Use the Run-3 recommended isolation and ID systematic prescriptions.** The MCP group has updated systematics strategies for Run-3 (revisiting dominant sources and how to treat them). Include those recommended uncertainty components in your systematics model. Indico

Sources

- Collected 5 sources
- Next job: Add auto-updating pipelines to CI/CD
  - Already in place for Markdown

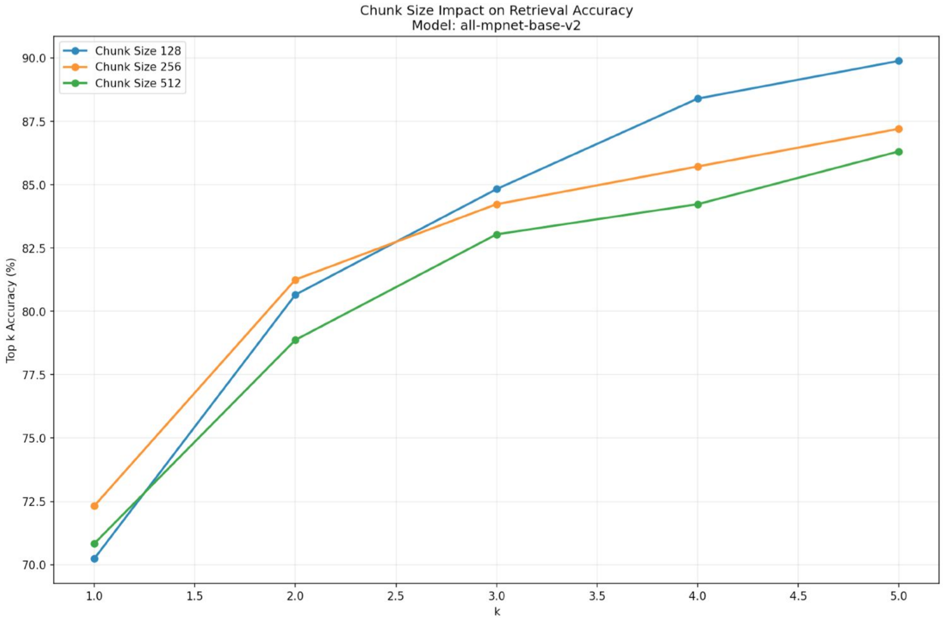
Name	Acquisition Method	# Parent chunks	# Child Chunks
Twiki	Direct access via CERN IT	148,707	661,153
CDS	Scraping	224,965	1,029,869
Indico	Scraping	6,491	32,969
ATLAS Talk	API	29,294	233,162
Markdown (Gitlab)	API (GraphQL)	45,838	201,681

Sources: Automate scraping

- Automated scrape/update pipeline using [Gitlab CI/CD runner]
    - Built on GraphQL
- Weekly execution, only scrape projects updated that week
- Filter out:
    - Projects containing blacklisted patterns
    - Projects with fewer than 5 markdown files, *not counting:*
        - CHANGELOG files
        - Files with fewer than 100 or more than 100k characters
    - Forks
    - Personal projects
    - Empty repos
    - Inactive projects (no commit in last 120 days)
    - Pull mirrors

Document chunking

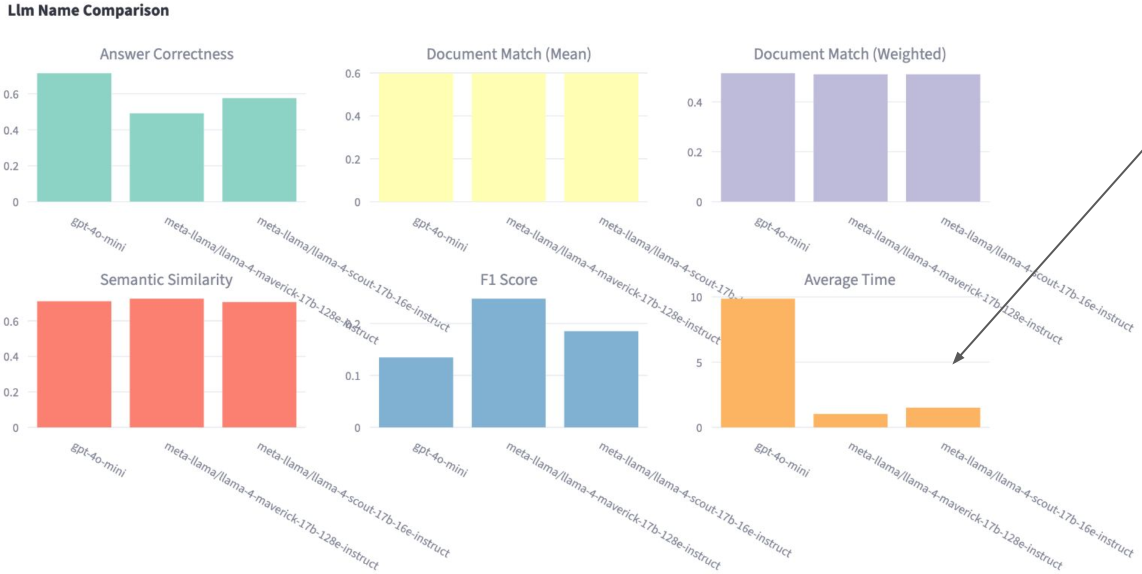
- **The idea:** Mapping a document to a single vector inherently causes information loss
- Break documents into smaller units to preserve more information
- Chunk size is a parameter that must be tuned
  - Alternate chunking strategies possible e.g. by headers
  - Tradeoff is no fixed chunk length



Postgres architecture optimisation

- Computing similarity scores is expensive (dot product of high dimensional vector)
- Look for algorithms to minimise computation during search
- Support for different vector DB search algorithms
  - IVFFlat (default)
  - HNSW
  - Exact search
- Postgres default parameters for IVFFlat and HNSW not optimal
  - Tuning required

Groq Hosted Models

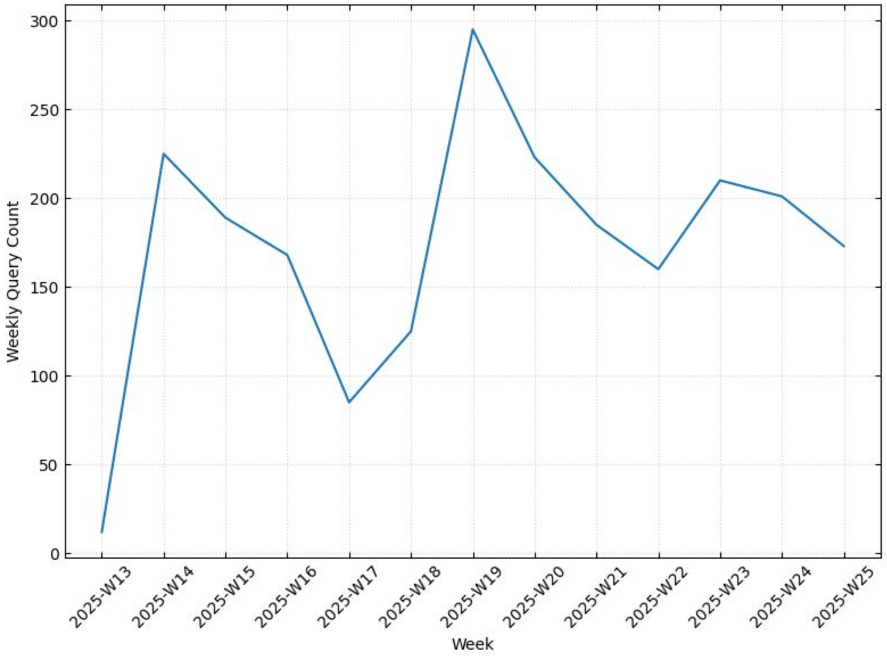


Order of magnitude lower latency

Benchmark: Increasing coverage

User questions

- Logging query-answer pairs for ~5 months, plan is:
  - Responses with positive feedback go into benchmark
  - Responses with negative feedback require manual review
- Currently not enough responses with feedback for this to be useful
  - Supplement benchmark with manually reviewed responses



Benchmark: Increasing coverage

LLM generated questions

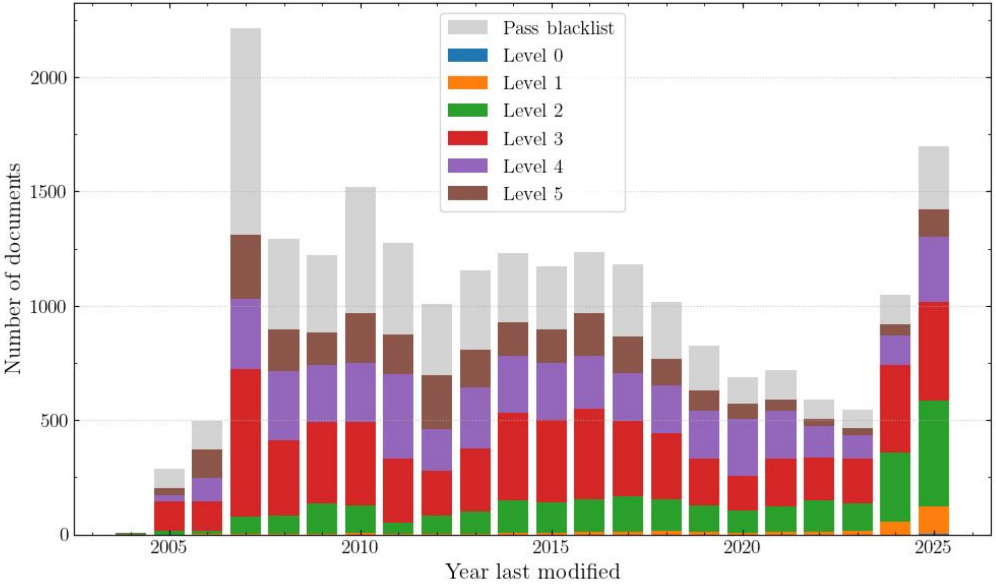
- Increase benchmark coverage with less personpower
- 750 important twikis linked from the top level ones
- Several validation stages using an LLM (see right)
- Questions passing these are reviewed manually

1. Relevant to ATLAS?
2. Outdated?
3. Appropriate technical depth?
4. Completely answerable from document?
5. Answer is factually accurate?
6. Answer represents the document content?

Ongoing Work: Garbage in, garbage out

- In principle, should make *all documentation* available for retrieval
- But retrieval is not 100% accurate, so in practice, you **only want to keep the most useful information**
- Current decision: Remove anything not updated in last 5 years
- Experimenting with methods to upweight important docs

Hierarchy of twikis defined by distance to top level



Ongoing Work: Getting more from embeddings

- Embedding models map text to vectors
- Can get further by **fine-tuning on HEP** documents
- Currently deployed: Transformer, finetuned on ATLAS talk Q&A pairs
- Recent progress: adding metadata to embeddings improves performance
- Finetuning on synthetic markdown Q&A pairs further improves markdown retrieval
    - Finetune a model for each source type

Credit: Sam Van Stroud (UCL)