chATLAS

An AI Assistant for the ATLAS Collaboration

DANIEL MURNANE, GABRIEL FACINI,
RUNZE LI & CARY RANDAZZO

AI4EIC WORKSHOP
NOVEMBER 30, 2023

BERKELEY LAB UCL Yale LOUISIANA TECH UNIVERSITY

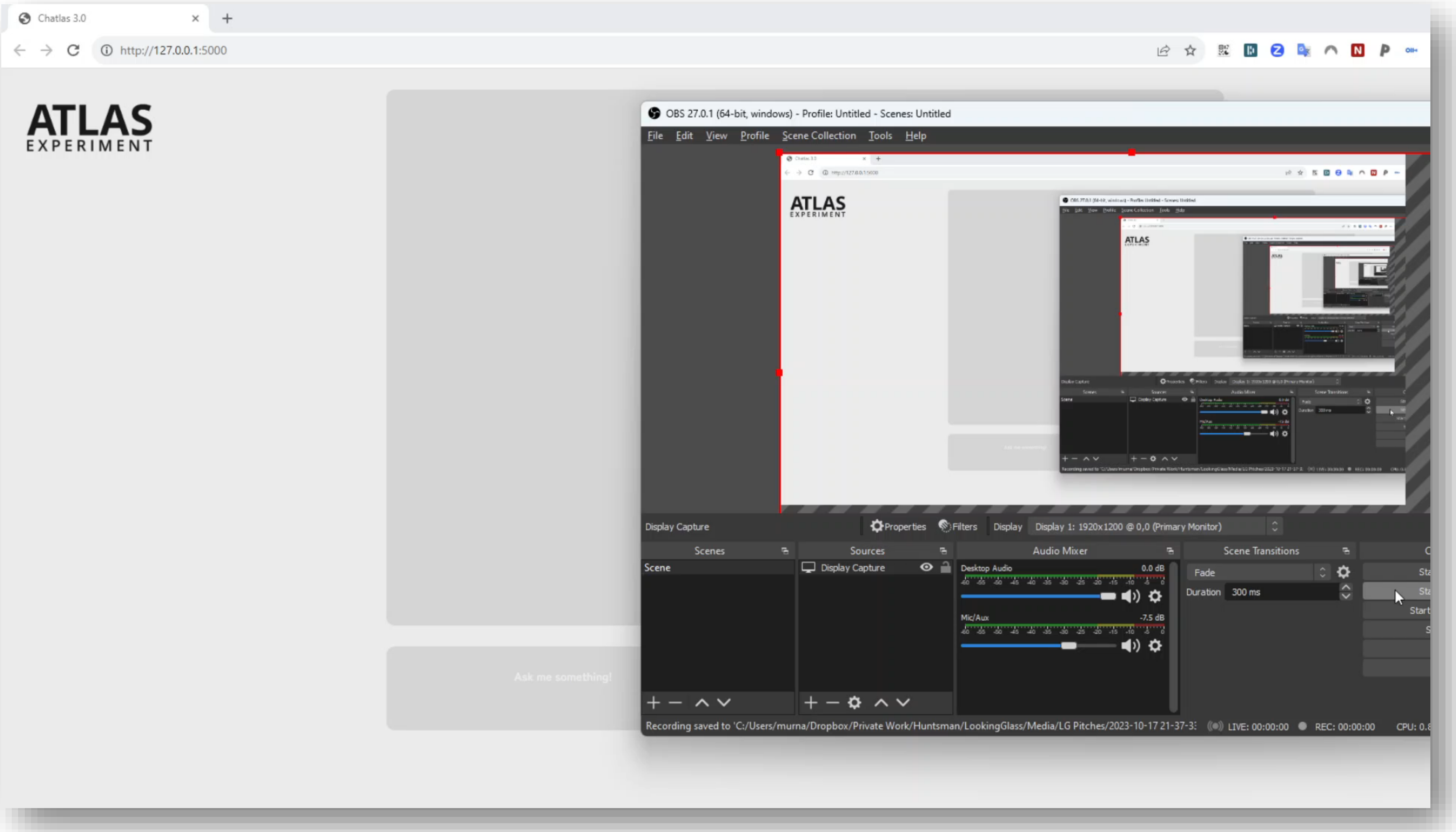
MOTIVATION

- Motivated to have a tool that could be a front-door for…
- Quickly parsing documentation and twiki
- Semantic search and availability of heterogeneous sources of ATLAS information
- Summarizing research
- Connecting the dots between different groups
- Debugging software
- Searching and summarizing JIRA and Glance information

HISTORY

- In April 2023, initial ATLAS ML Forum meeting to discuss usage of ChatGPT and Github Copilot within ATLAS
- In June 2023, presentations on several ongoing works to use LLMs within ATLAS
  - ATLAS-GPT: Daniel Murnane
  - ChATLAS: Gabriel Facini
  - Google Bard + ATLAS: Kaushik De
  - Analysis Description Language + GPT: Gokhan Unel
- Decision made to converge ATLAS-GPT and ChATLAS and create an official prototype
- Fortnightly developer meetings kicked off in August 2023
- Currently approx. seven part-time contributors
- Launched ATLAS-public demo November 16 https://chatlas-flask-chatlas.app.cern.ch/

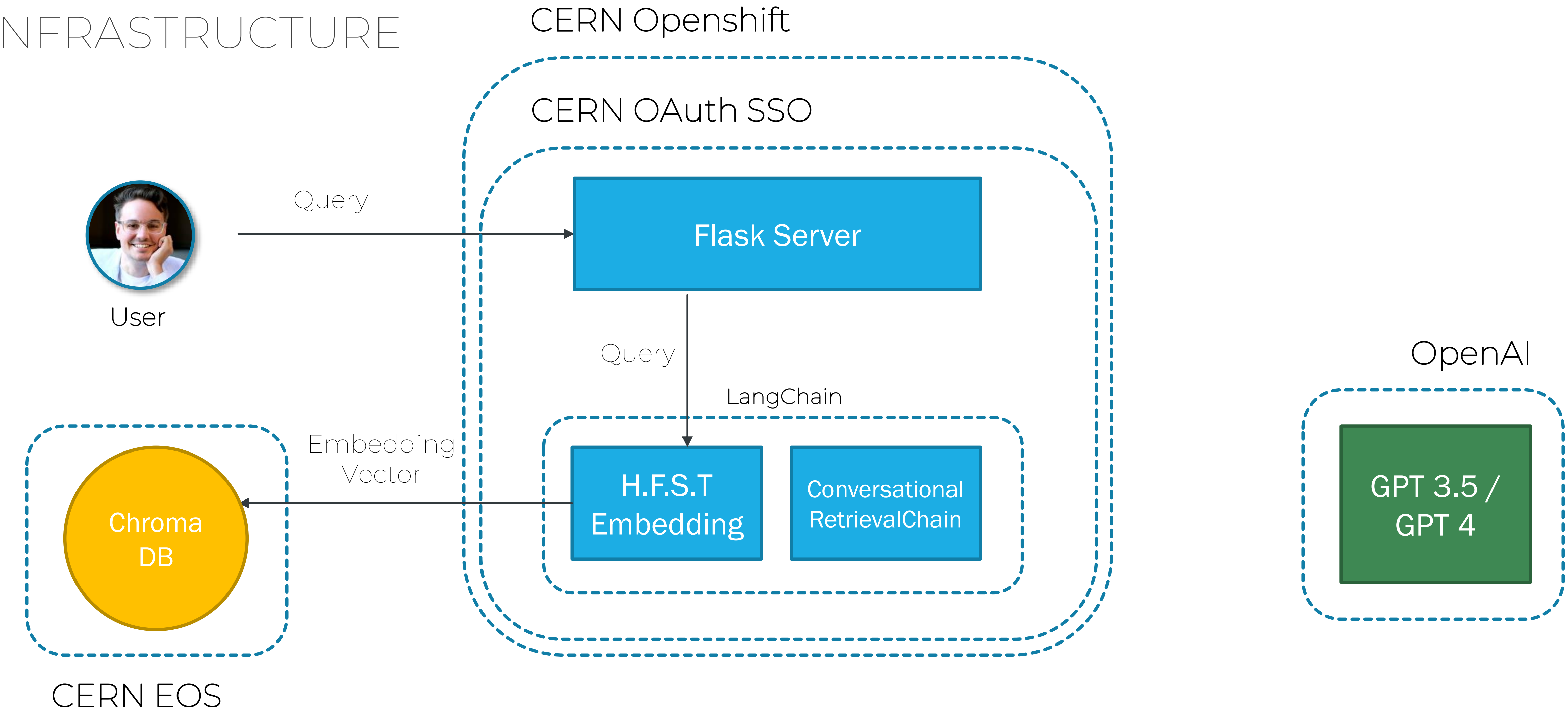
CURRENT STATUS: DEMO



Recording saved to 'C:/Users/murna/Dropbox/Private Work/Huntsman/LookingGlass/Media/LG Pitches/2023-10-17 21-37-3

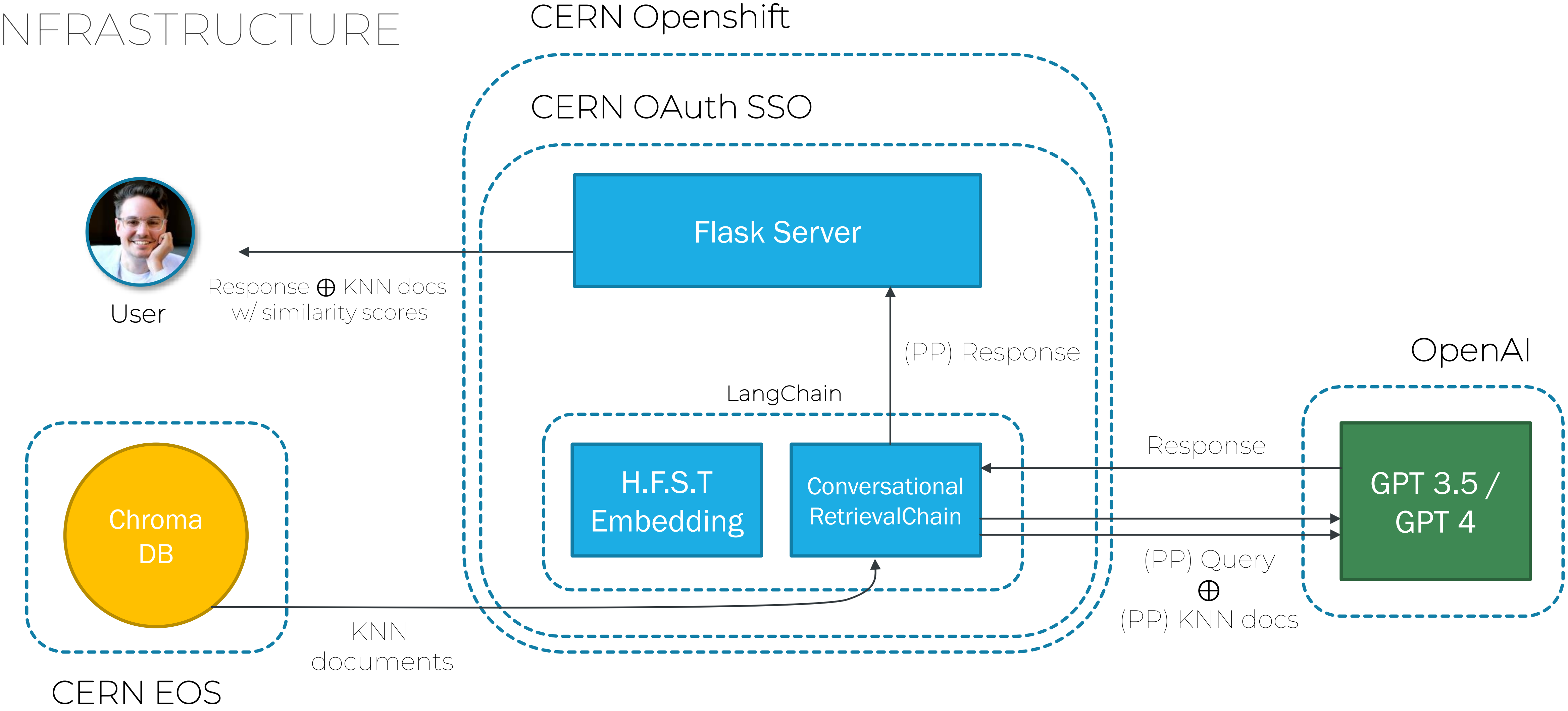
CURRENT STATUS: INFRASTRUCTURE

(PP) = (Possibly Processed)



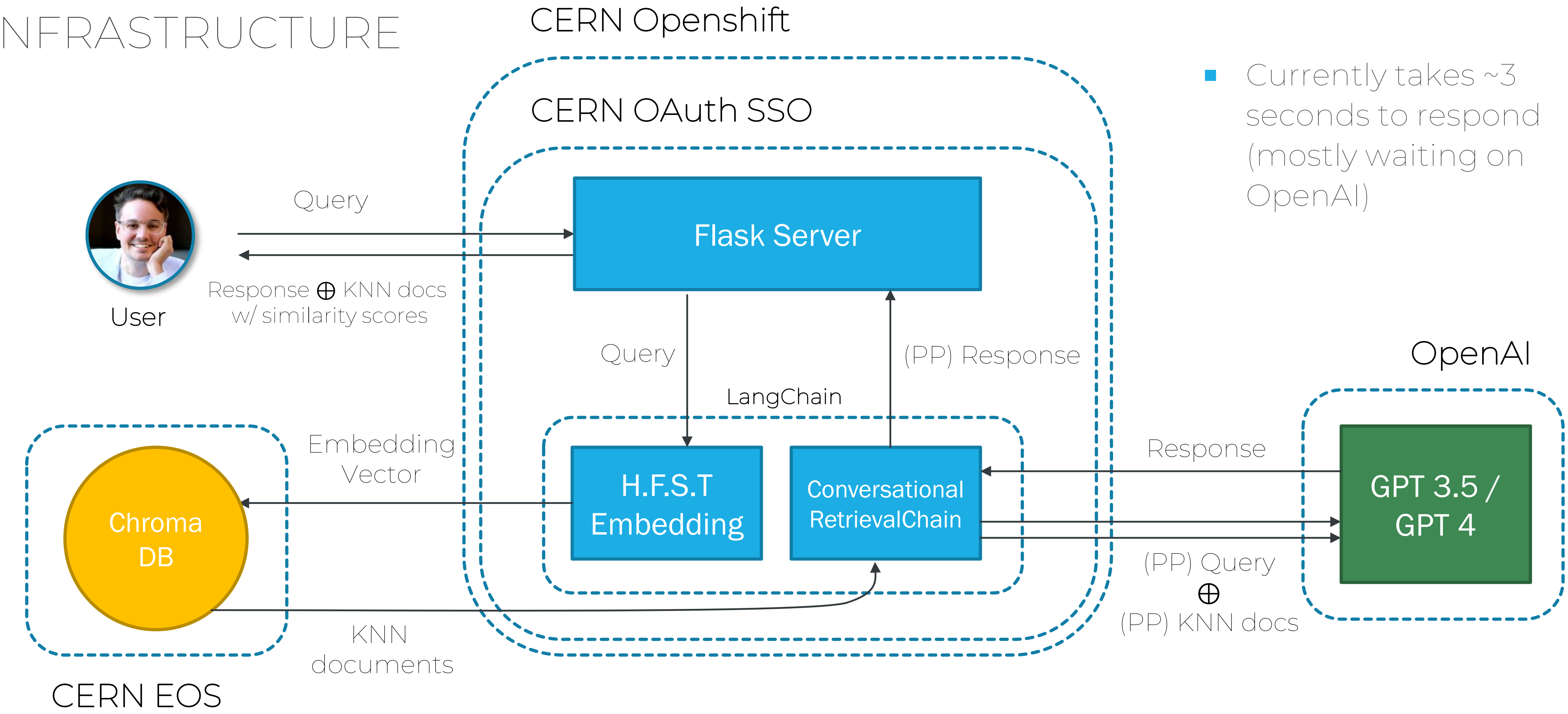
CURRENT STATUS: INFRASTRUCTURE

(PP) = (Possibly Processed)



CURRENT STATUS: INFRASTRUCTURE

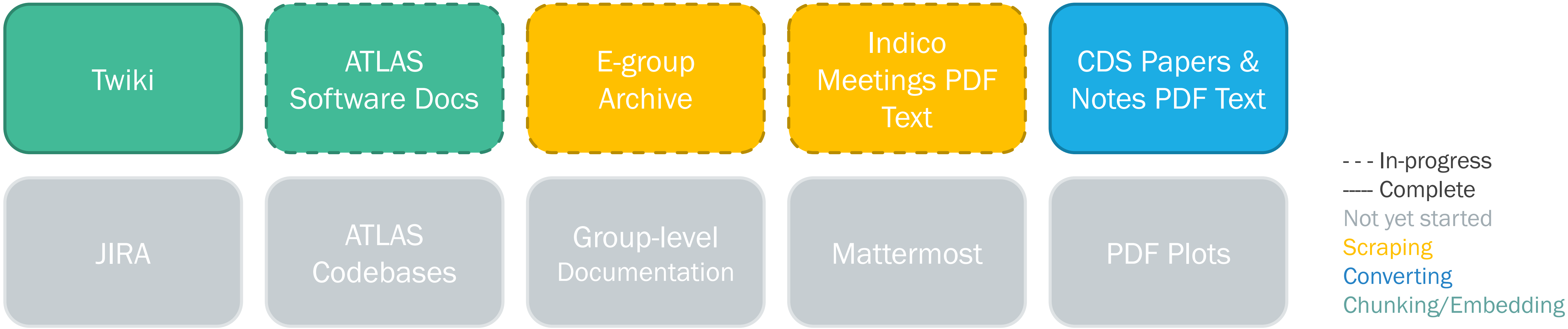
(PP) = (Possibly Processed)



- Currently takes ~3 seconds to respond (mostly waiting on OpenAI)

CURRENT STATUS: DATASETS

- Diagram of all the possible ATLAS datasets and how many we have in the DB
- Add a box around those that are *almost* in the DB
- Stage the progress of each (solid/dashed in-progress/complete, scraped, converted, chunked+embedded)



Twiki

ATLAS Software Docs

E-group Archive

Indico Meetings PDF Text

CDS Papers & Notes PDF Text

JIRA

ATLAS Codebases

Group-level Documentation

Mattermost

PDF Plots

- - - In-progress
----- Complete
Not yet started
Scraping
Converting
Chunking/Embedding

DETAILS: SCRAPING

ATLAS Twiki

- Start with set of "Starting URLs"
- Recursively visit included links
- Find all headers, and visit content below
- Append metadata of twiki (parent structure, date revised, etc.)

CDS

- Discover whether the CDS paper has a Gitlab latex repo
- If latex exists, pull from repo and (planned) convert to markdown
- (Planned) Use **unstructured** library to parse markdown
- If latex *does not* exist, use **nougat** library to read PDF (including equations) into markdown

Indico

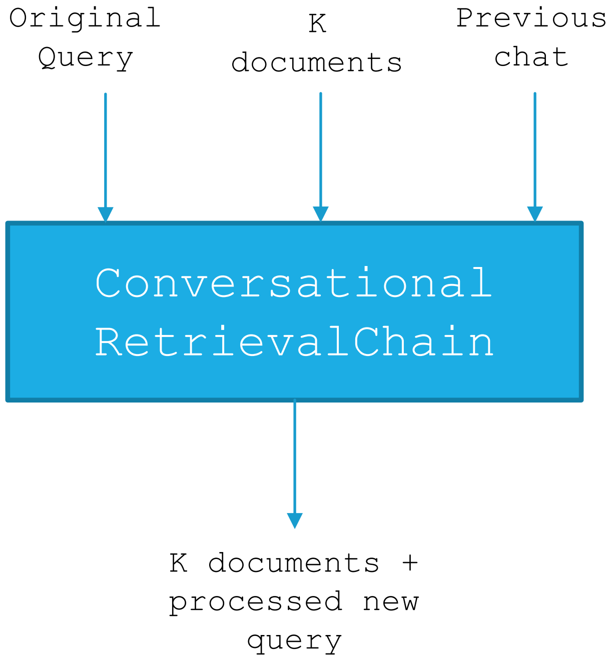
- Load event list
- Scrape timetable contents (date, title, speaker, etc.)
- (Planned) Pull PDF slide decks and minutes
- (Planned) Parse in the same way as in CDS

DETAILS: CHUNKING

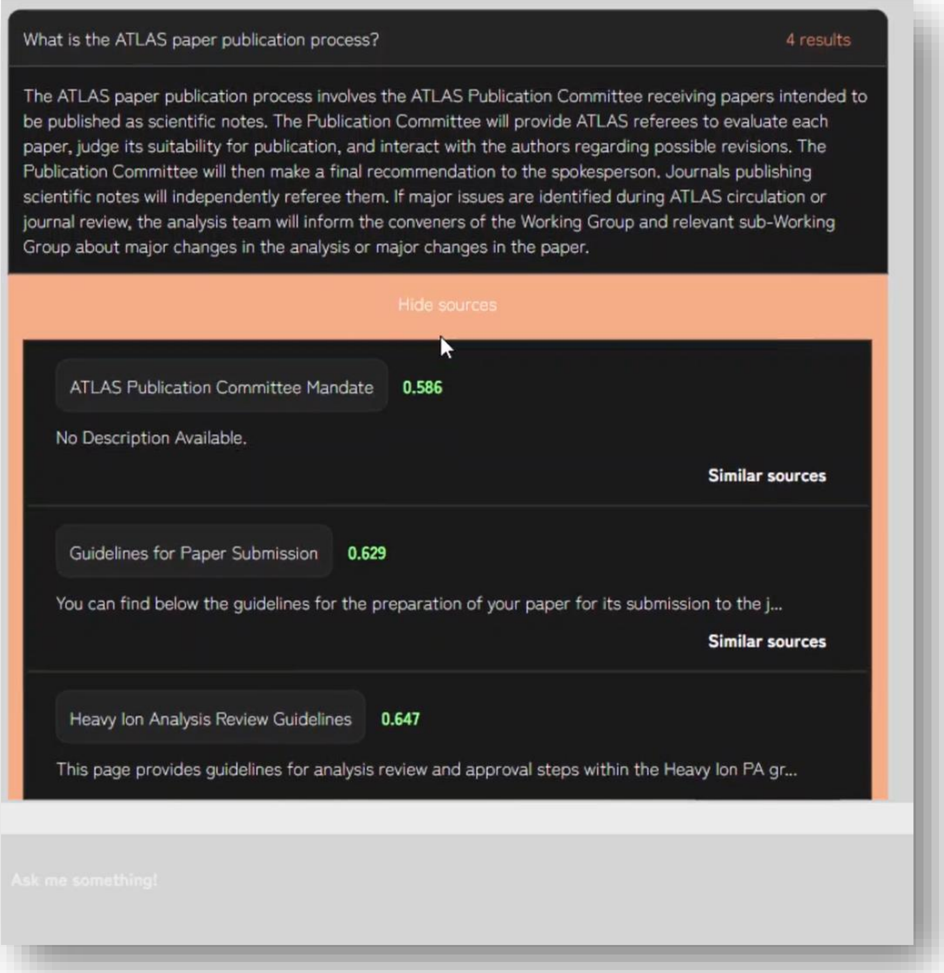
- (Current) Loop through HTML and Markdown heading sections
- If section exceeds 510 tokens, split with **SentenceTransformersTokenTextSplitter**
- Pass chunk through **HuggingFace's sentence-transformers/all-MiniLM-L6-v2** model
- (Planned) Use built-in **unstructured** library to identify chunks
- Insert chunk into Chroma database, with metadata of file URL, twiki name

DETAILS: RETRIEVAL

- All handled internally by
  ```
  qa = ConversationalRetrievalChain.from_llm(    llm=model,
  retriever=db.as_retriever(),    memory=memory,    verbose=False,)
  ```
- LLM Model is GPT-3.5 from OpenAI API, retriever is default Chroma which contains the embedding model, memory is a buffer that retains all previous chat information
- Implicit is that the model aggregates all K-documents with a prompt to produce a **new question** based on the original question and the K-documents


Original Query
K documents
Previous chat
ConversationalRetrievalChain
K documents + processed new query

DETAILS: CHAT INTERFACE



What is the ATLAS paper publication process?

The ATLAS paper publication process involves the ATLAS Publication Committee receiving papers intended to be published as scientific notes. The Publication Committee will provide ATLAS referees to evaluate each paper, judge its suitability for publication, and interact with the authors regarding possible revisions. The Publication Committee will then make a final recommendation to the spokesperson. Journals publishing scientific notes will independently referee them. If major issues are identified during ATLAS circulation or journal review, the analysis team will inform the conveners of the Working Group and relevant sub-Working Group about major changes in the analysis or major changes in the paper.

ATLAS Publication Committee Mandate 0.586

No Description Available.

Guidelines for Paper Submission 0.629

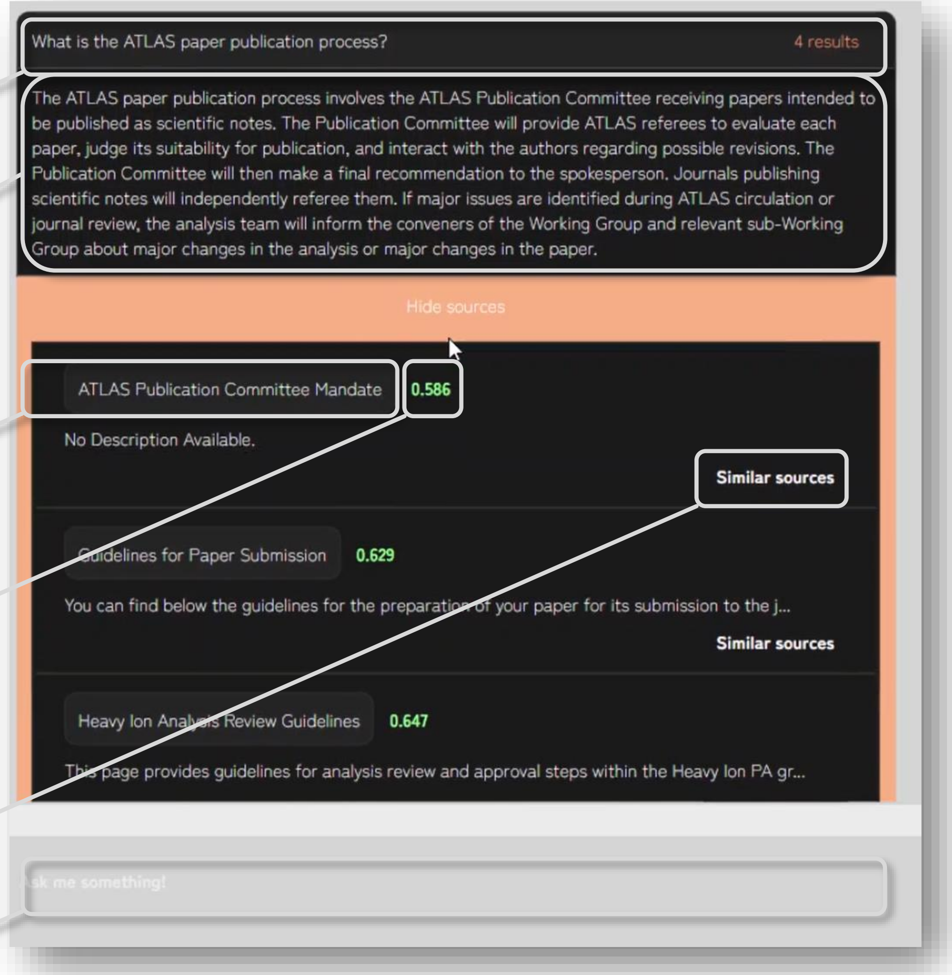
You can find below the guidelines for the preparation of your paper for its submission to the j...

Heavy Ion Analysis Review Guidelines 0.647

This page provides guidelines for analysis review and approval steps within the Heavy Ion PA gr...

DETAILS: CHAT INTERFACE

- Version 1.0 contains everything needed to answer a query
- Can cite the top sources used in the response
- Has a quick search for similar sources



What is the ATLAS paper publication process? 4 results

The ATLAS paper publication process involves the ATLAS Publication Committee receiving papers intended to be published as scientific notes. The Publication Committee will provide ATLAS referees to evaluate each paper, judge its suitability for publication, and interact with the authors regarding possible revisions. The Publication Committee will then make a final recommendation to the spokesperson. Journals publishing scientific notes will independently referee them. If major issues are identified during ATLAS circulation or journal review, the analysis team will inform the conveners of the Working Group and relevant sub-Working Group about major changes in the analysis or major changes in the paper.

ATLAS Publication Committee Mandate 0.586

No Description Available.

Guidelines for Paper Submission 0.629

You can find below the guidelines for the preparation of your paper for its submission to the j...

Heavy Ion Analysis Review Guidelines 0.647

This page provides guidelines for analysis review and approval steps within the Heavy Ion PA gr...

Input query

LLM response

Source title

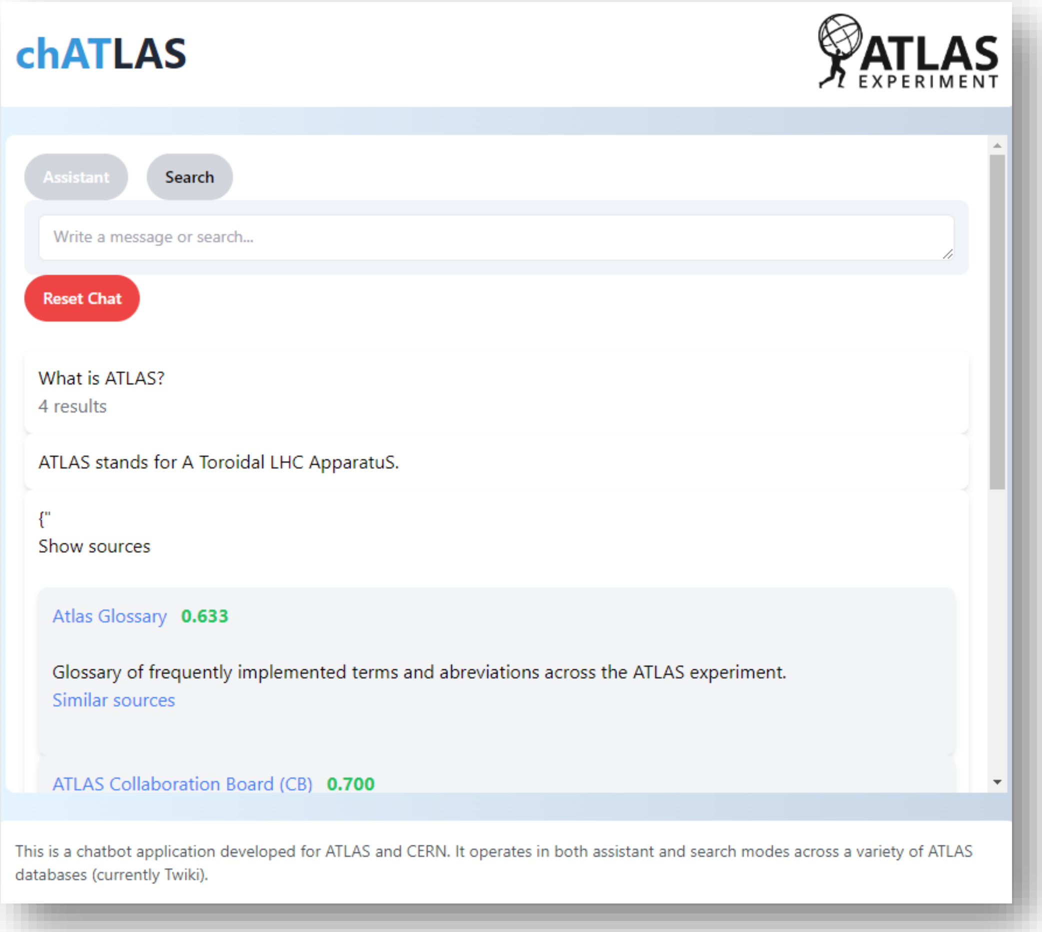
Similarity score

Instant similarity search

Next query entry

DETAILS: CHAT INTERFACE

- Version 1.0 contains everything needed to answer a query
- Can cite the top sources used in the response
- Has a quick search for similar sources
- Experimenting with a V2.0 appearance that is lighter, and has a dedicated Assistant mode and Search mode



What is ATLAS?
4 results

ATLAS stands for A Toroidal LHC ApparatuS.

{"
Show sources

Atlas Glossary **0.633**

Glossary of frequently implemented terms and abreviations across the ATLAS experiment.
Similar sources

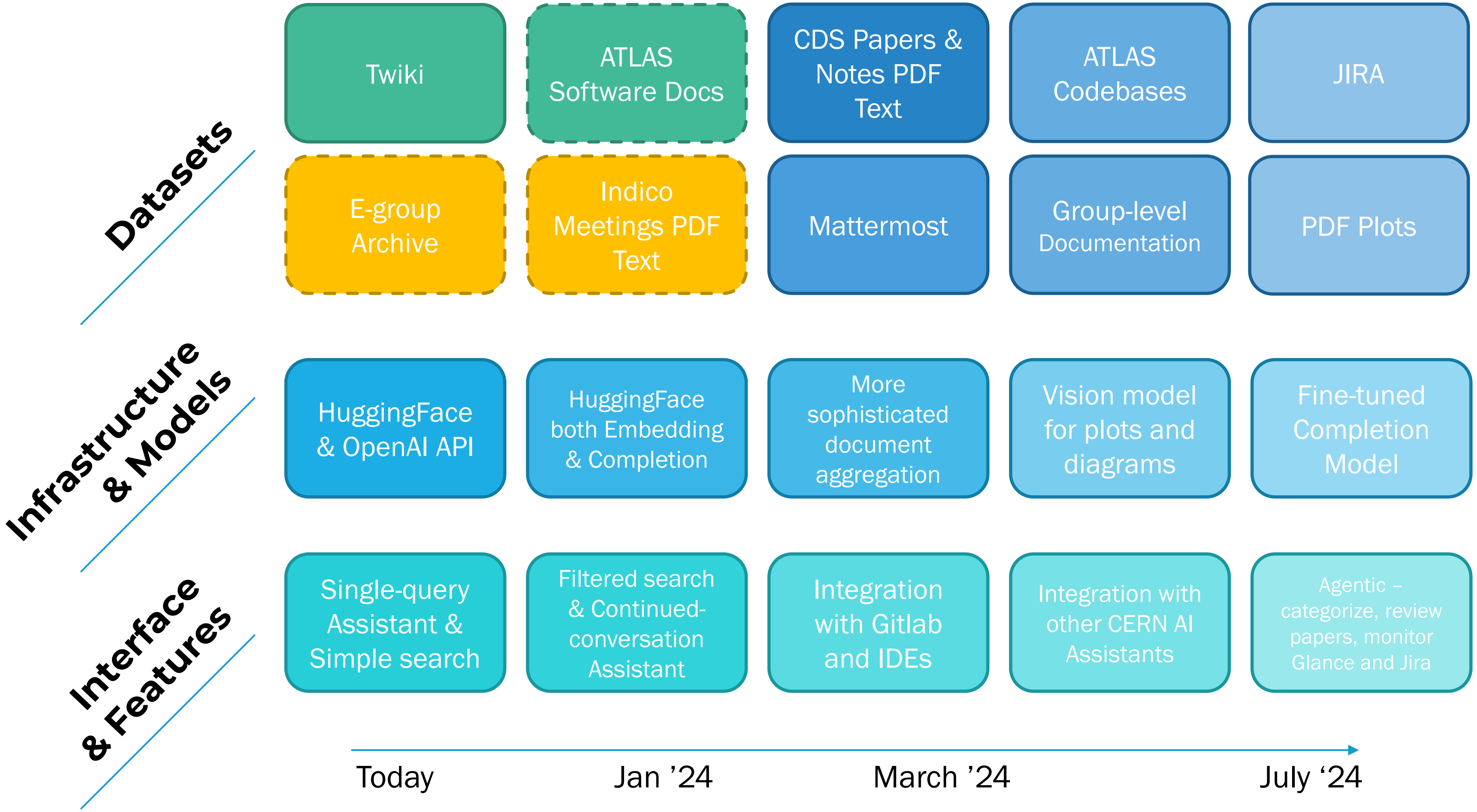
ATLAS Collaboration Board (CB) **0.700**

This is a chatbot application developed for ATLAS and CERN. It operates in both assistant and search modes across a variety of ATLAS databases (currently Twiki).

SHORTCOMINGS AND HURDLES

- Getting the data! Highly heterogeneous file types, many behind authorisation walls, many stale or inaccurate, many requiring high levels of post-processing
- Community solutions could go a long way: Ensure that any experiment/collaboration databases are easily accessible and exportable. All websites should live in a git repo. All publications should be submitted and saved as latex, and compiled separately. All discussion forums should have anonymisation options. This would have saved O(1 year) of data wrangling
- Hallucination is still a very real problem [https://www.arxiv-vanity.com/papers/2311.04348/]
- A high quality AI assistant probably requires fine tuning, which is an expensive task (less in gpu-hours, more in expert-hours)
- Open-source solutions for UI are not particularly flexible. A tool built by+for the scientific community would be **very** useful! Open-source solutions for backend (retrieval, document aggregation) are perfectly fine.
- Codebase integration: experiment codebases are huge, not so well-commented, and non-obvious how to chunk. Perhaps an automated commenting algorithm as a pre-process step?

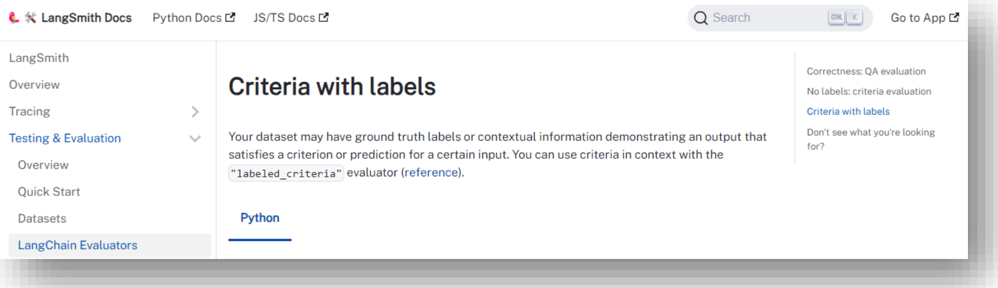
ROADMAP



Datasets	Twiki	ATLAS Software Docs	CDS Papers & Notes PDF Text	ATLAS Codebases	JIRA
	E-group Archive	Indico Meetings PDF Text	Mattermost	Group-level Documentation	PDF Plots
Infrastructure & Models	HuggingFace & OpenAI API	HuggingFace both Embedding & Completion	More sophisticated document aggregation	Vision model for plots and diagrams	Fine-tuned Completion Model
Interface & Features	Single-query Assistant & Simple search	Filtered search & Continued-conversation Assistant	Integration with Gitlab and IDEs	Integration with other CERN AI Assistants	Agentic – categorize, review papers, monitor Glance and Jira

OPEN QUESTIONS & CONCLUSION

- How to avoid hallucinations? Probably "GPT-5 / Q* / AGI" will make this hurdle irrelevant
- How to best "censor" politically incorrect responses (e.g. which analysis team is the best?)
- How to **measure** the quality of responses – LangSmith AI-assisted evaluators?
- What is the best dataset to gather for fine-tuning?
- How to anonymize email threads and discussion forums?



Criteria with labels

Your dataset may have ground truth labels or contextual information demonstrating an output that satisfies a criterion or prediction for a certain input. You can use criteria in context with the `"labeled_criteria"` evaluator (reference).

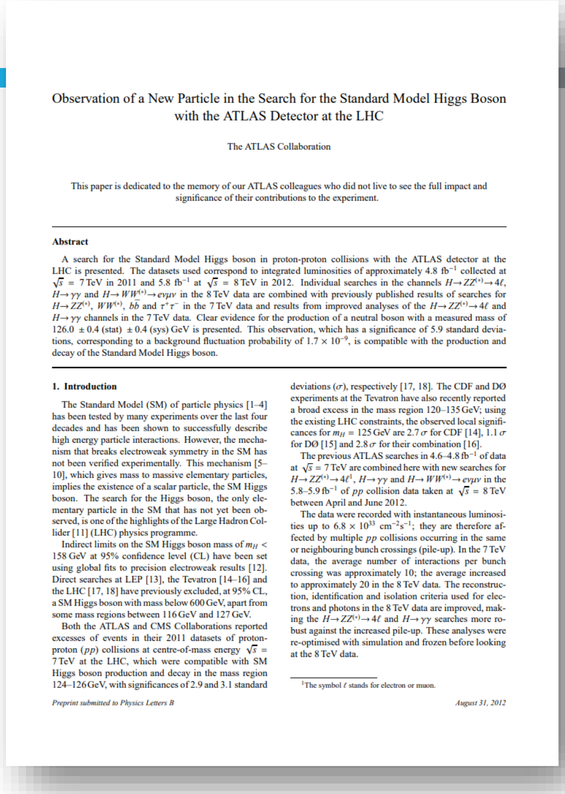
We are having a lot of fun building this thing from scratch, but if there was an open-source scientific community framework for AI Assistants, it would be even more fun!

BACKUP

RETRIEVAL-AUGMENTED GENERATION

RETRIEVAL-AUGMENTED GENERATION (RAG)

1. Convert document to txt



Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$, $H \rightarrow \gamma\gamma$ and $H \rightarrow WW^{(*)} \rightarrow e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \rightarrow ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$ and $H \rightarrow \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of 126.0 ± 0.4 (stat) ± 0.4 (sys) GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

---

1. Introduction

The Standard Model (SM) of particle physics [1–4] has been tested by many experiments over the last four decades and has been shown to successfully describe high energy particle interactions. However, the mechanism that breaks electroweak symmetry in the SM has not been verified experimentally. This mechanism [5–10], which gives mass to massive elementary particles, implies the existence of a scalar particle, the SM Higgs boson. The search for the Higgs boson, the only elementary particle in the SM that has not yet been observed, is one of the highlights of the Large Hadron Collider [11] (LHC) physics programme.

Indirect limits on the SM Higgs boson mass of $m_H < 158$ GeV at 95% confidence level (CL) have been set using global fits to precision electroweak results [12]. Direct searches at LEP [13], the Tevatron [14–16] and the LHC [17, 18] have previously excluded, at 95% CL, a SM Higgs boson with mass below 600 GeV, apart from some mass regions between 116 GeV and 127 GeV.

Both the ATLAS and CMS Collaborations reported excesses of events in their 2011 datasets of proton-proton (pp) collisions at centre-of-mass energy $\sqrt{s} = 7$ TeV at the LHC, which were compatible with SM Higgs boson production and decay in the mass region 124–126 GeV, with significances of 2.9 and 3.1 standard deviations (σ), respectively [17, 18]. The CDF and DØ experiments at the Tevatron have also recently reported a broad excess in the mass region 120–135 GeV; using the existing LHC constraints, the observed local significances for $m_H = 125$ GeV are 2.7 σ for CDF [14], 1.1 σ for DØ [15] and 2.8 σ for their combination [16].

The previous ATLAS searches in 4.6–4.8 fb^{-1} of data at $\sqrt{s} = 7$ TeV are combined here with new searches for $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$[1], $H \rightarrow \gamma\gamma$ and $H \rightarrow WW^{(*)} \rightarrow e\nu\mu\nu$ in the 5.8–5.9 fb^{-1} of pp collision data taken at $\sqrt{s} = 8$ TeV between April and June 2012.

The data were recorded with instantaneous luminosities up to 6.8×10^{33} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$; they are therefore affected by multiple pp collisions occurring in the same or neighbouring bunch crossings (pile-up). In the 7 TeV data, the average number of interactions per bunch crossing was approximately 10; the average increased to approximately 20 in the 8 TeV data. The reconstruction, identification and isolation criteria used for electrons and photons in the 8 TeV data are improved, making the $H \rightarrow ZZ^{(*)} \rightarrow 4\ell$ and $H \rightarrow \gamma\gamma$ searches more robust against the increased pile-up. These analyses were re-optimised with simulation and frozen before looking at the 8 TeV data.

---

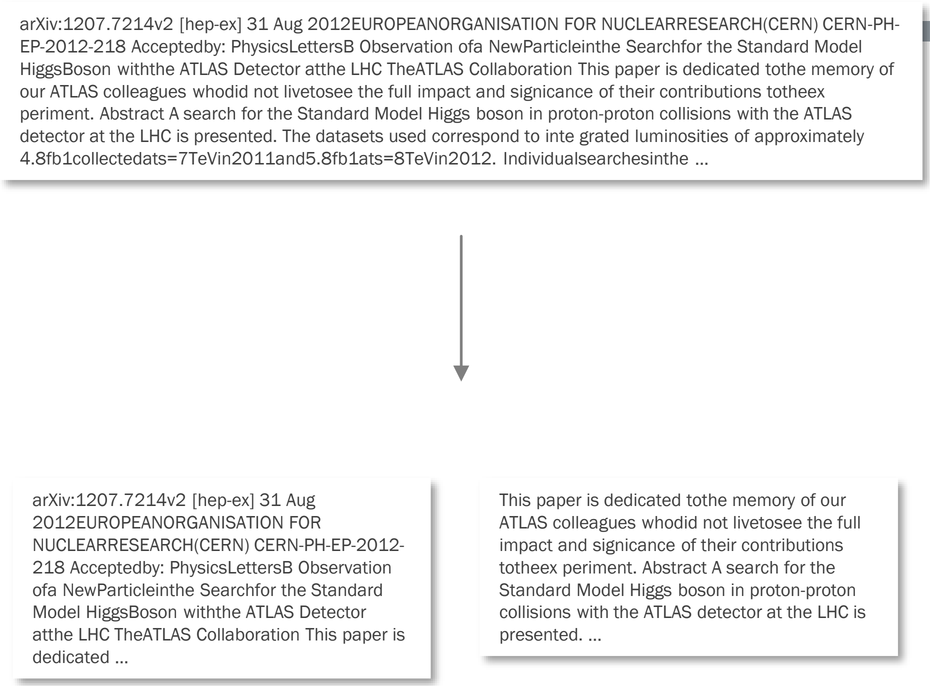
[1] The symbol ℓ stands for electron or muon.

Preprint submitted to Physics Letters B

August 31, 2012

```
arXiv:1207.7214v2 [hep-ex] 31 Aug 2012EUROPEANORGANISATION FOR NUCLEARRESEARCH(CERN) CERN-PH-EP-2012-218 Acceptedby: PhysicsLettersB Observation ofa NewParticleinthe Searchfor the Standard Model HiggsBoson withthe ATLAS Detector atthe LHC TheATLAS Collaboration This paper is dedicated tothe memory of our ATLAS colleagues whodid not livetosee the full impact and signicance of their contributions totheex periment. Abstract A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to inte grated luminosities of approximately 4.8fb1collectedats=7TeVin2011and5.8fb1ats=8TeVin2012. Individualsearchesinthe …
```

RETRIEVAL-AUGMENTED GENERATION (RAG)



arXiv:1207.7214v2 [hep-ex] 31 Aug 2012EUROPEANORGANISATION FOR NUCLEARRESEARCH(CERN) CERN-PH-EP-2012-218 Acceptedby: PhysicsLettersB Observation ofa NewParticleinthe Searchfor the Standard Model HiggsBoson withthe ATLAS Detector atthe LHC TheATLAS Collaboration This paper is dedicated tothe memory of our ATLAS colleagues whodid not livetosee the full impact and signicance of their contributions totheex periment. Abstract A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to inte grated luminosities of approximately 4.8fb1collectedats=7TeVin2011and5.8fb1ats=8TeVin2012. Individualsearchesinthe ...

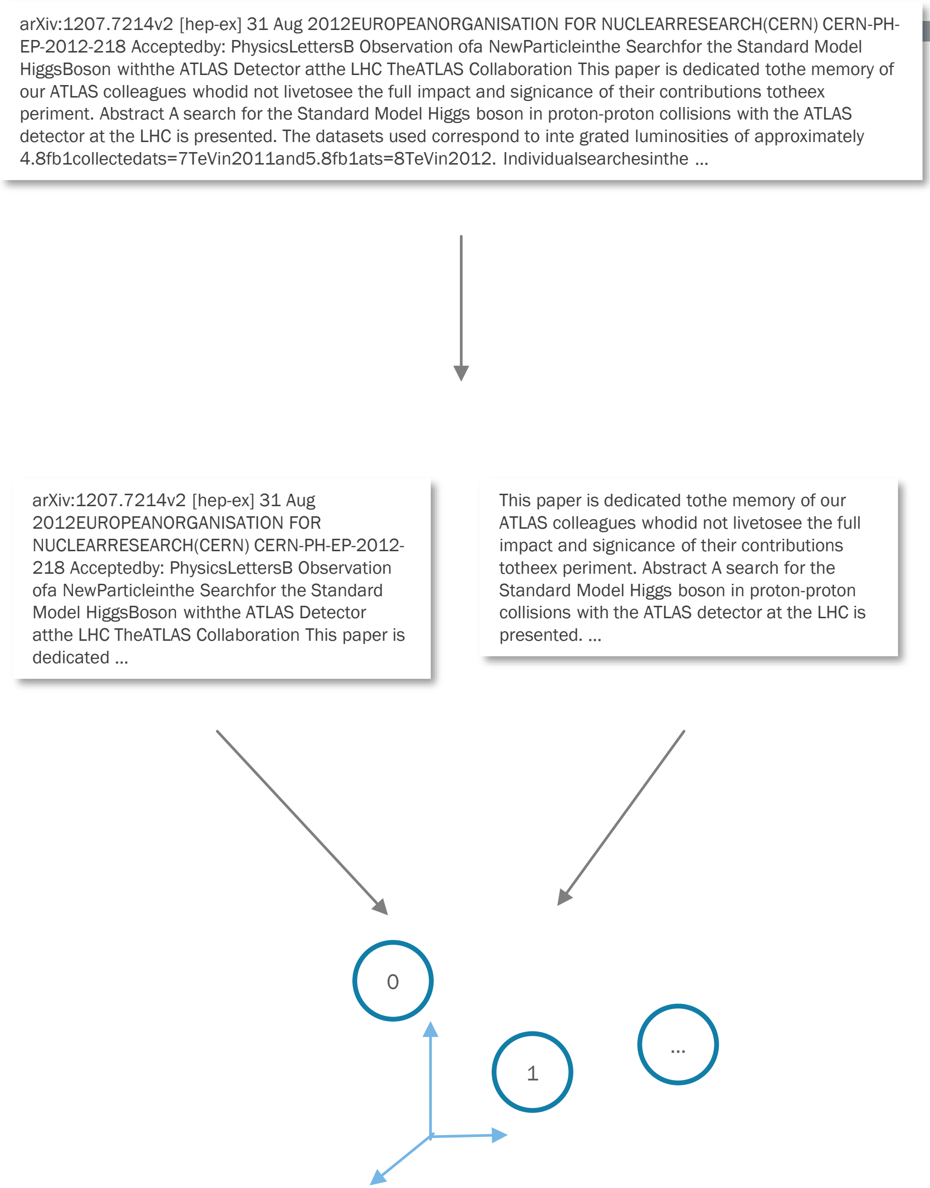
arXiv:1207.7214v2 [hep-ex] 31 Aug 2012EUROPEANORGANISATION FOR NUCLEARRESEARCH(CERN) CERN-PH-EP-2012-218 Acceptedby: PhysicsLettersB Observation ofa NewParticleinthe Searchfor the Standard Model HiggsBoson withthe ATLAS Detector atthe LHC TheATLAS Collaboration This paper is dedicated ...

This paper is dedicated tothe memory of our ATLAS colleagues whodid not livetosee the full impact and signicance of their contributions totheex periment. Abstract A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. ...

1. Convert document to txt
2. Split into K-word chunks

RETRIEVAL-AUGMENTED GENERATION (RAG)

1. Convert document to txt
2. Split into K-word chunks
3. Embed each chunk with HuggingFace SentenceTransformer model into 1536-dimensional vector space



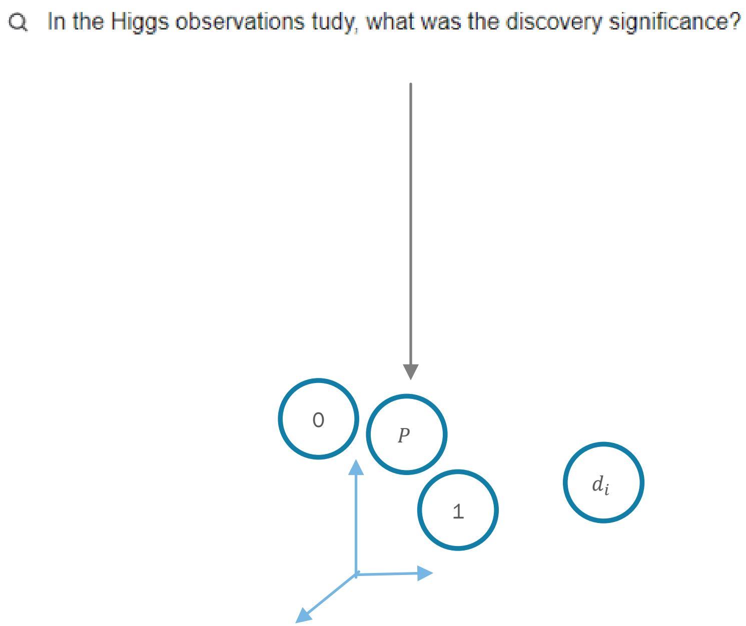
arXiv:1207.7214v2 [hep-ex] 31 Aug 2012EUROPEANORGANISATION FOR NUCLEARRESEARCH(CERN) CERN-PH-EP-2012-218 Acceptedby: PhysicsLettersB Observation ofa NewParticleinthe Searchfor the Standard Model HiggsBoson withthe ATLAS Detector atthe LHC TheATLAS Collaboration This paper is dedicated tothe memory of our ATLAS colleagues whodid not livetosee the full impact and significance of their contributions totheex periment. Abstract A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to inte grated luminosities of approximately 4.8fb1collectedats=7TeVin2011and5.8fb1ats=8TeVin2012. Individualsearchesinthe ...

arXiv:1207.7214v2 [hep-ex] 31 Aug 2012EUROPEANORGANISATION FOR NUCLEARRESEARCH(CERN) CERN-PH-EP-2012-218 Acceptedby: PhysicsLettersB Observation ofa NewParticleinthe Searchfor the Standard Model HiggsBoson withthe ATLAS Detector atthe LHC TheATLAS Collaboration This paper is dedicated ...

This paper is dedicated tothe memory of our ATLAS colleagues whodid not livetosee the full impact and significance of their contributions totheex periment. Abstract A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. ...

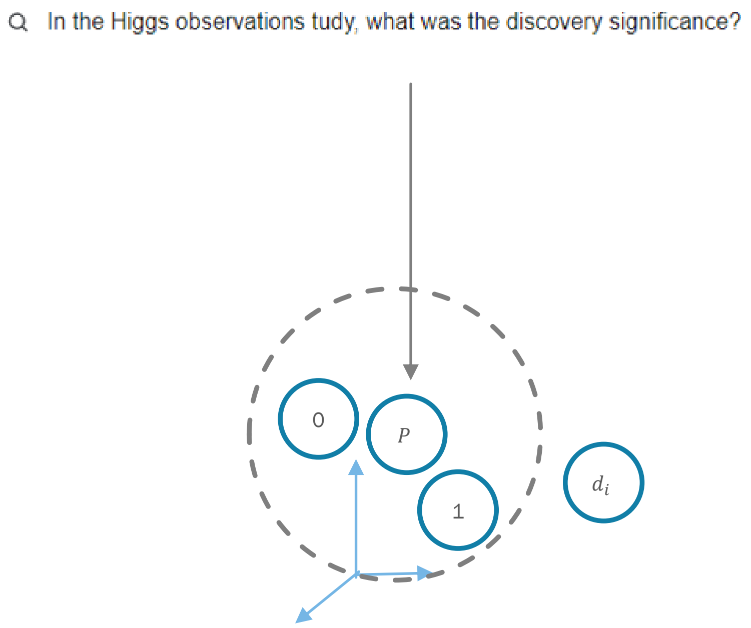
HOW IT WORKS

1. Convert document to txt
2. Split into K-word chunks
3. Embed each chunk with HuggingFace SentenceTransformer model into 1536-dimensional vector space
4. Embed prompt $\boldsymbol{P}$ in this space


In the Higgs observations tudy, what was the discovery significance?
0
P
1
d_i

HOW IT WORKS

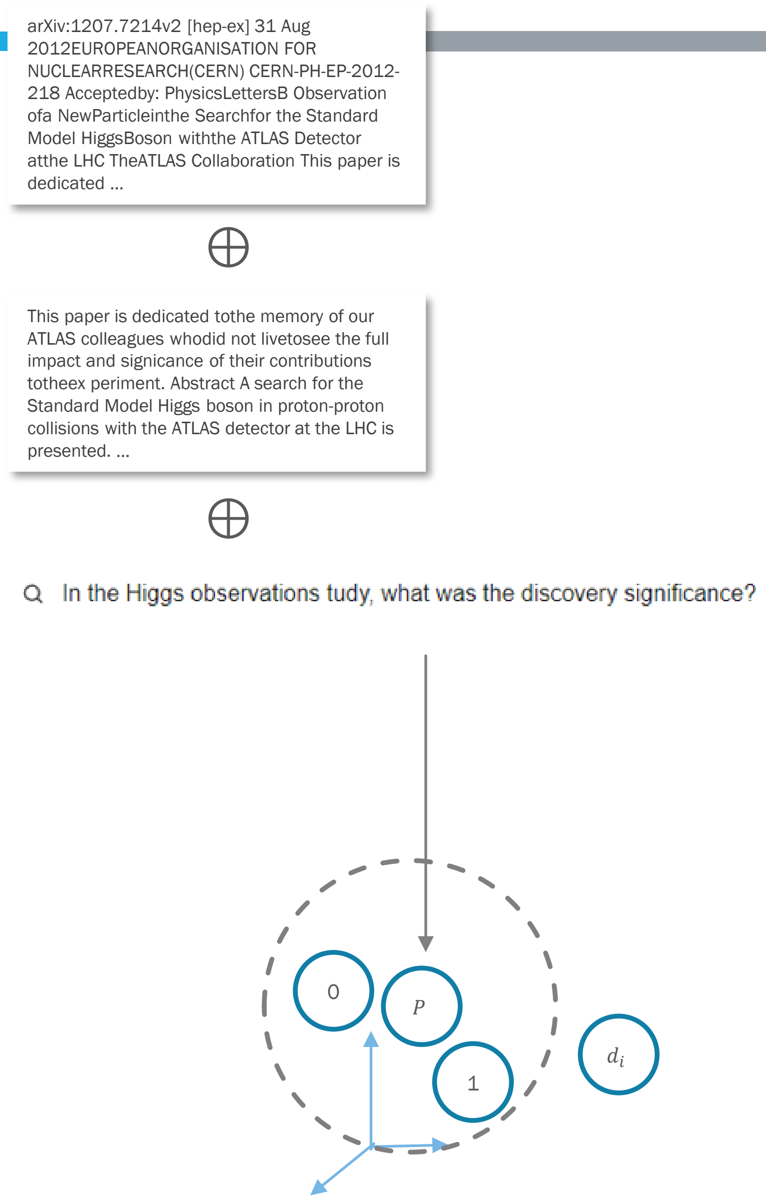
1. Convert document to txt
2. Split into K-word chunks
3. Embed each chunk with HuggingFace SentenceTransformer model into 1536-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search (KNN, $K \leq 10$, $P \cdot d_i \leq 0.5$)



In the Higgs observations tudy, what was the discovery significance?

HOW IT WORKS

1. Convert document to txt
2. Split into K chunks
3. Embed each chunk with HuggingFace SentenceTransformer model into 1536-dimensional vector space
4. Embed prompt P in this space
5. Nearest neighbour search (KNN, $K \leq 10$, $\mathrm{P} \cdot d_i \leq 0.5$)
6. Append all matching sections to prompt and query GPT-4



arXiv:1207.7214v2 [hep-ex] 31 Aug 2012EUROPEANORGANISATION FOR NUCLEARRESEARCH(CERN) CERN-PH-EP-2012-218 Acceptedby: PhysicsLettersB Observation ofa NewParticleinthe Searchfor the Standard Model HiggsBoson withthe ATLAS Detector atthe LHC TheATLAS Collaboration This paper is dedicated ...

⊕

This paper is dedicated tothe memory of our ATLAS colleagues whodid not livetosee the full impact and significance of their contributions totheex periment. Abstract A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. ...

⊕

In the Higgs observations tudy, what was the discovery significance?