chATLAS

An AI Assistant for the ATLAS Collaboration

DANIEL MURNANE, GABRIEL FACINI,
RUNZE LI, DANIELE DEL SANTO & CARY RANDAZZO

1st Large Language Models in Physics Symposium
February 21, 2024

chATLAS Overview

- Motivation
- History of chATLAS
- Data Gathering
- Infrastructure
- Deployment Result: Chat Interface and Demo
- Roadmap
- Updates, Challenges, and Summary

MOTIVATION

- 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/

Data Gathering: How Scraping Began

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

Data Gathering: Volume Measurements

Twiki

ATLAS Software Docs

E-group Archive

Indico Meetings

PDF Plots

Mattermost

Jira

ATLAS Codebases

Group level Docs

CDS Papers & Notes

Data Gathering: Volume Measurements and Tool Selection

Twiki

Over 2000 ATLAS Twiki Topics

TWiki > AtlasProtected Web > AtlasPhysics (2024-02-08, MariaSmizanska)

ATLAS

ATLAS Physics Activity

Physics coordinators: Monica Dunford, Fabio Cerutti

Analysis - Getting started and key links

Getting set-up
Data checklist for physics analyses - Starting page for GRLs, data and MC
AMG info - Useful FAQ, info on PHYS/PHYSLITE, software tutorials, who to ask for help

Finding your samples
Data and Monte Carlo Datasets - Starting point for all data/MC information
Derivations
MC Sample Request Procedure

Trigger and Object calibration
CP analysis Kick-off work - CP work for your analysis
Run 3 CP recommendations - Status and links
Run 2 CP recommendations - Status and links
Lowest unprescaled triggers
Luminosity

High priority CP work
Tracking | EGamma | Muon | Tau | JetETmiss | Flavour tagging | Simulation | PMG | Physics validation

Hepdata and public data
ATLAS Hepdata recommendations - helpful tips and scrips to get started

General links
How the ATLAS collaboration works - All helpful ATLAS links in one place!
Guidelines to produce event displays - create event displays
Run 2 Discrepancies and excesses

Paper writing links and tips

Approval of Results - links approval procedure and policies
PubComm main page - Contains all instructions, latex templates, etc
Conference deadlines for papers and CONF notes - timelines for major conferences
ATLAS figure style recommendations (Jun 2014)
ATLAS rounding rules (Mar 2014)
Physics Office - policy documents, glance-related instructions, egroups naming rules
Glance Phase 0 instructions - including who is on what egroup

Policy Documents
ATLAS Policy Documents - includes policy on theses, job talks, etc

Events

ATLAS physics workshops - Workshops and guidelines on away meetings
ATLAS posters at LHCC meeting, November 27th, 2023

Physics Organisation

Physics Coordination Membership
Operation Task Planner (OTP)
SCAB - Instructions on SCAB
List of qualification tasks
Compact list of talks at main meetings (indicomb)

Task forces

Combined Performance (CP) Groups

Egamma	L. Aperio Bella, K. Lohwasser
Flavour Tagging	F.A. Di Bello, D. Guest
Inner Tracking	C. Grefe, N. Calace
JetEtMiss	R. Camacho Toro, F. Balli
Muon	R. Nikolaidou, S. Angelidakis
Tau	S.M. Farrington, A. De Maria

Physics Analysis (PA) Groups

B Physics & Light States	A. Cerri, S. Turchikhin
Exotics	T. Vazquez Schroeder, D. Hayden
Heavy Ions	A. Angerami, Q. Hu
Higgs	N. Berger, P. Francavilla
Higgs & Diboson Searches	E. Brost, A. Cortes Gonzalez
Physics Modelling	D. Hirschbuehl, A. Cueto Gomez
Standard Model	P. Sommer, S. Camarda
Supersymmetry	J. Montejo Berlingen, S. Alderweireld
Top	A. Knue, N.A. Asbah
Upgrade Physics	A. Schwartzman, H. De La Torre Per

List of PA & CP subgroups and subgroup convenors
Physics and CP Group Liaisons
ATLAS Appointments Database
List of CMS conveners

Other Groups related to PC

Analysis Model Group	E. Kourlitis, G. Watts
Boosted Higgs bbcc tagger	A. Coccaro, F. Filthaut
Derivation Coordination and Production Team	E. Torro Pastor
Forward proton performance	R. Staszewski, S.E. Clawson
Global particle flow task force	M. Hodgkinson, M. Swiatlowski
Isolation and Fake Forum	N. Bruscino, F. Alonso
Join EFT and Interpretation group	S. Kortner, T. Dado, H. Mildner
Luminosity WG (data preparation) (info)	V.S. Lang, E. Torrence
Machine Learning Forum	W.H. Hopkins, D.T. Murnane
MC Production	E.M. Lobodzinska, Y.L. Liu
Physics Office	G. Navarro

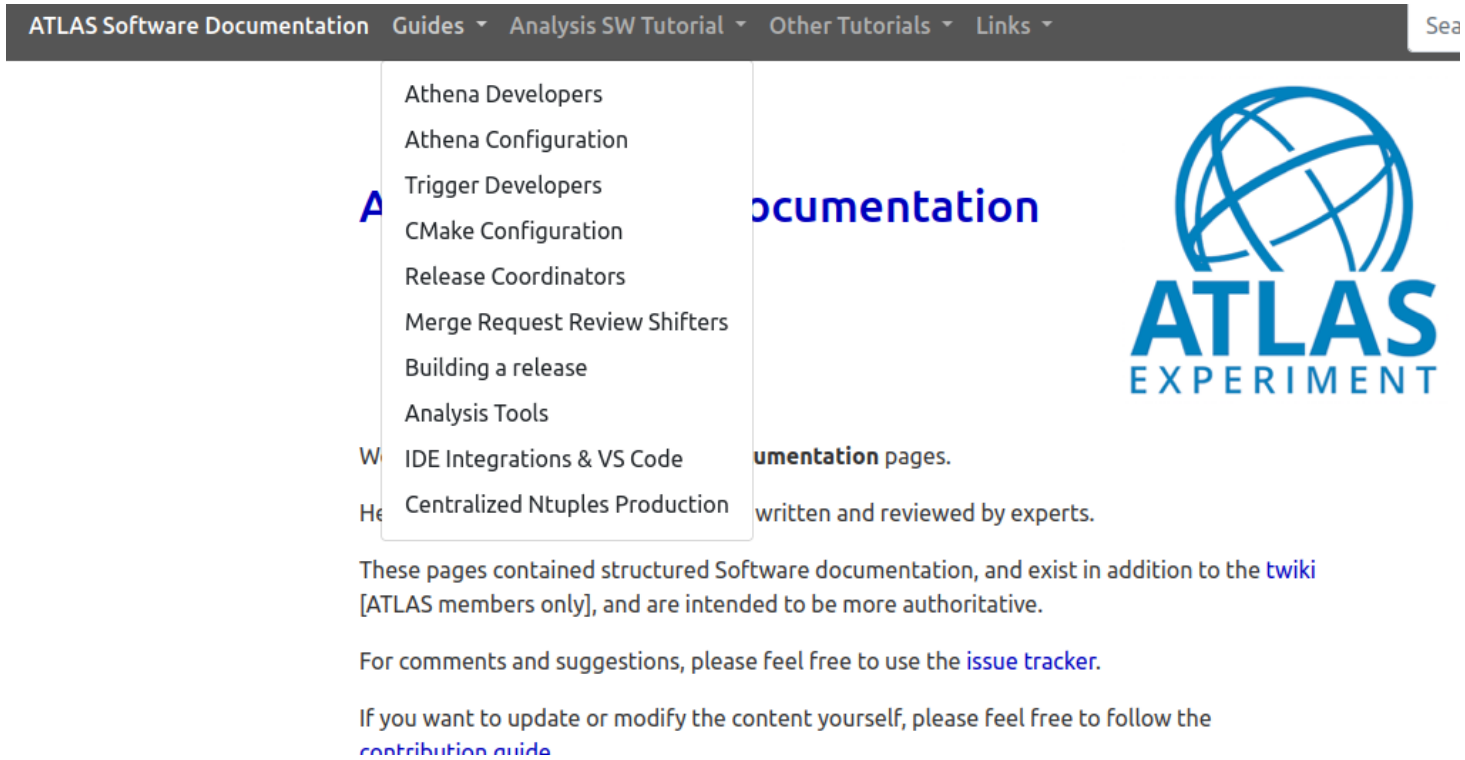
Using:
BeautifulSoup and
auth-get-sso-cookie

Data Gathering: Volume Measurements and Tool Selection

ATLAS Software Docs

Using:
BeautifulSoup and
auth-get-sso-cookie

Hundreds of ATLAS Software Docs



Data Gathering: Volume Measurements and Tool Selection

E-group Archive

Atlas forums

Archive	E-group name	Description
Category : Computing Documentation and Announcements (5)		
CERN Computing Announcements	hn-atlas-cernCompAnnounce	This forum will contain announcements of changes, outages and othe
Documentation and Communication	hn-atlas-docAndCommunication	This forum covers general announcements concerning Documentatio
Grid Announcements	hn-atlas-gridAnnounce	This forum contains announcements related to the Grid infrastructure
Releases and Distribution Kit Announcements	hn-atlas-releaseKitAnnounce	This Forum announces the Release Plans and New Releases or patcl Do not reply to the postings, send comments or reports on problems t
Software Developers Announcements	hn-atlas-SWDevelopersAnnounce	Announcements, which are intended for developers of ATLAS offline s Do not reply to postings in this Forum, send comments or reports on p
Category : Computing Offline Software (28)		
Architecture Team: Core Software Architecture and Design	hn-atlas-SWArchitecture	A forum to follow and participate to the work of the ATLAS Architecture
Athena-ROOT access	hn-atlas-athena-ROOT-access	Developer's discussion of progress in accessing Athena objects from
Atlantis Event Display	hn-atlas-AtlantisDisplay	Forum for help requests, suggestions, discussions and comments reg
Atlfast Support	hn-atlas-atlfast-support	Support for Atlas fast simulation users
Atlfast Working Group	hn-atlas-atlfast2-newsdev-val	This forum is aimed for the development, validation and testing of the
Bugs	hn-atlas-Prelimbugs	Atlas Preliminary/unconfirmed Bugs, Problems, Frustrations, Fixes
Bytestream Initiative	hn-atlas-bytestream-initiative	A finite-lifetime forum to coordinate bytestream infrastructure developr
Digitization Developers	hn-atlas-digitization-developers	This forum is intended for coordinating and discussing digitization soft
Fatras News, Development and Validation	hn-atlas-fatras	This forum is aimed at the developers of Fatras such as the users. In
General Offline Help	hn-atlas-offlineSWHelp	Requests for guidance in writing or using software. Comments on reco Please report release specific errors and problems to the Forum "Rele
Generator Validation	hn-atlas-generator-validation	Common forum for MC generator responsibles in ATLAS
Forum for VP1 support	hn-atlas-vp1-help	Forum for VP1(Virtual Point 1) support
New Job Configuration	hn-atlas-NewJobConfiguration	in this forum we will discuss the migration to Configurables, JobPrope
Offline Commissioning	hn-atlas-offline-commissioning	Offline Commissioning: For discussion about the offline software used
Offline SW Development Discussions	hn-atlas-offlineSWDevelopment	This group is for the Athena and Package developers to discuss desig
Persistency Help	hn-atlas-persistencyHelp	This forum is a general help list for reading and writing data. It is mea
Physics and Software Validation	hn-atlas-physics-software-validation	Common forum for discussions and announcements concerning Soft
PileUp	hn-atlas-pileup	To discuss and improve pile-up simulation, reconstruction and perform
Reconstruction Bug Monitor	hn-atlas-reco-bug-monitor	Forum for developers who wish to receive mail whenever a Reconstru
Reconstruction Integration	hn-atlas-recoIntegration	This forum is the main discussion channel for the Reconstruction Integ Although the RIG has a number of defined members, this forum is ope For software development question not particularly related to reconstr
Releases and Distribution Kit Problems	hn-atlas-releaseKitProblem	This Forum is for requests for assistance with problems running the sc

8912 topics x (1-50+ messages) for largest egroup
~ tens of thousands of messages to date

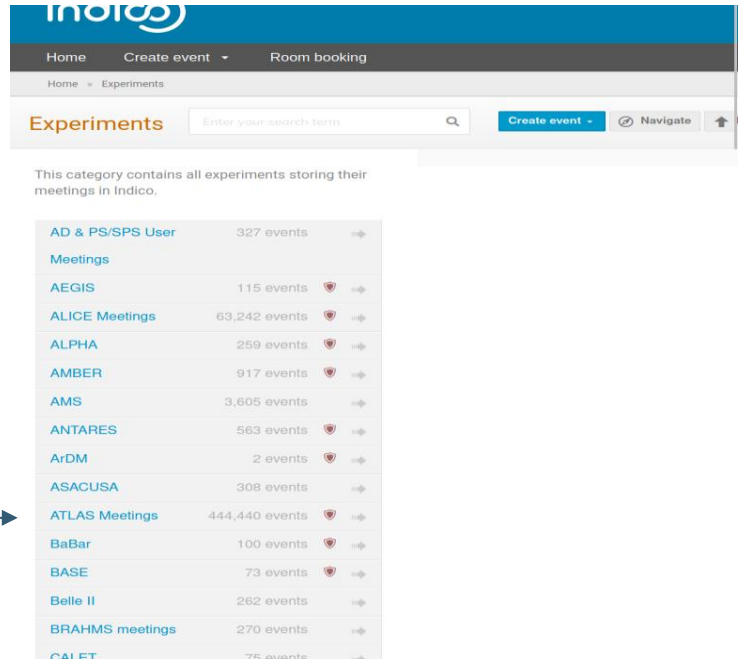
Using: BeautifulSoup and Selenium

Name	Size
hn-atlas-lar-electronic-calibration.txt	304.5 MB
hn-atlas-triggerReleaseValidation.txt	84.6 MB
hn-atlas-TDAQCommissioning.txt	81.1 MB
hn-atlas-tile-performance.txt	68.8 MB
hn-atlas-dist-analysis-help.txt	39.2 MB
hn-atlas-muonSWBugMonitor.txt	26.6 MB
hn-atlas-data-quality-operations.txt	18.1 MB
hn-atlas-offlineSWHelp.txt	17.4 MB
hn-atlas-PATHelp.txt	17.2 MB
hn-atlas-SITInternal.txt	8.2 MB
hn-atlas-exotics-wg.txt	6.9 MB
hn-atlas-jet-etmiss-wg.txt	6.5 MB

Data Gathering: Volume Measurements and Tool Selection

Indico Meetings

Using:
Nougat and Marker



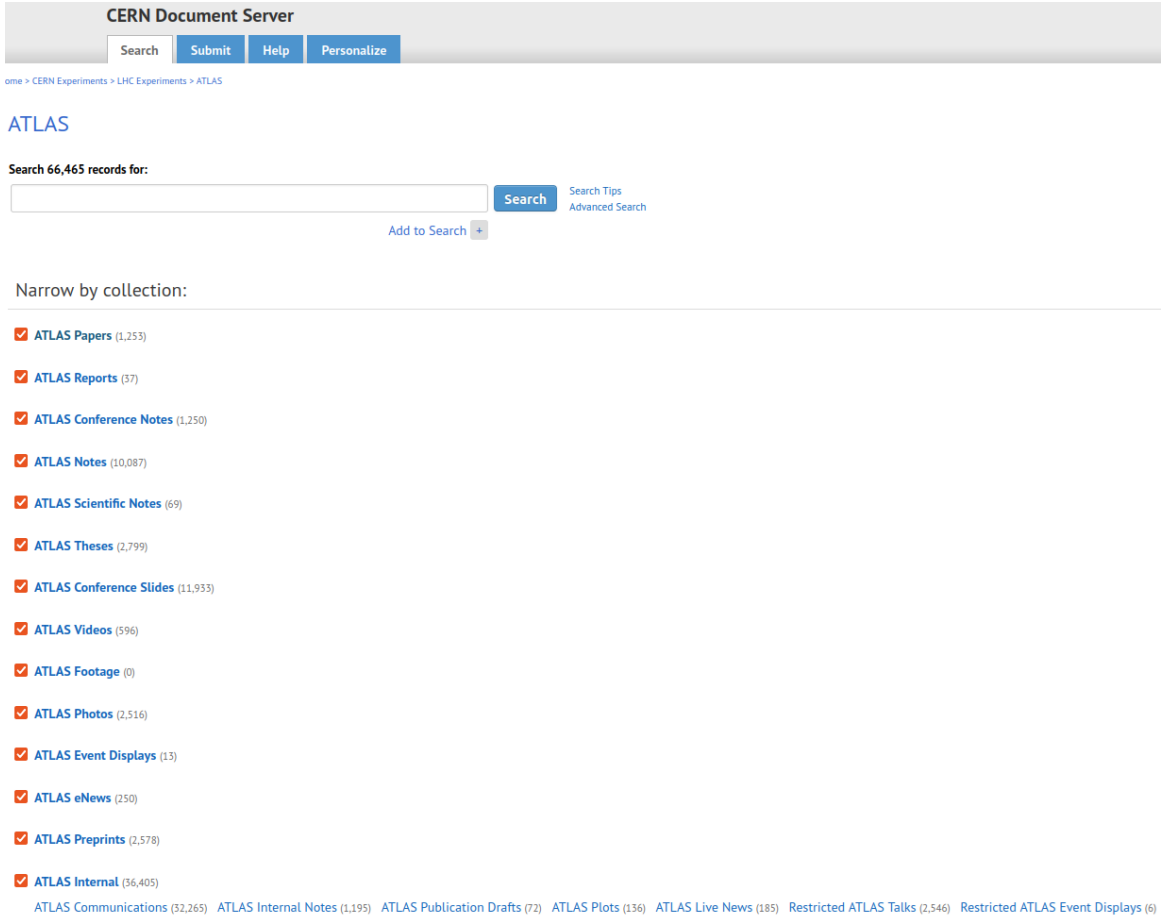
440,440 ATLAS Indico Meeting Events

Data Gathering: Volume Measurements and Tool Selection

PDF Plots	Mattermost	Jira	ATLAS Codebases	Group level Docs
			Measurements Pending - very large	Using: Pending further experiments

Data Gathering: Volume Measurements and Tool Selection

CDS Papers & Notes



Using:
Nougat and Marker

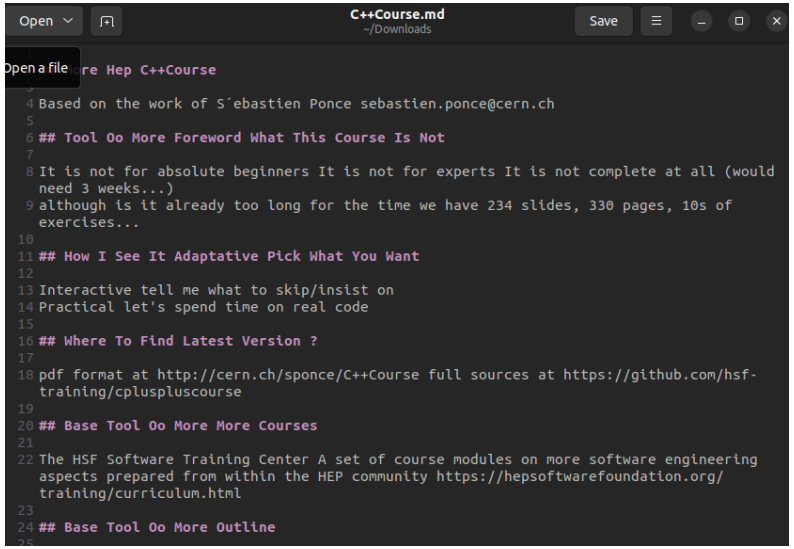
Total: 66,465 records found

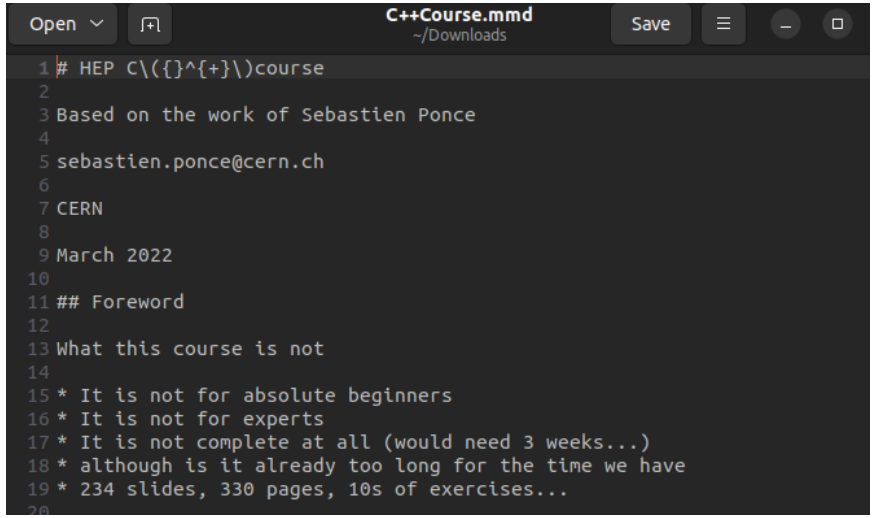
36,405 are Internal including communications, notes, etc.

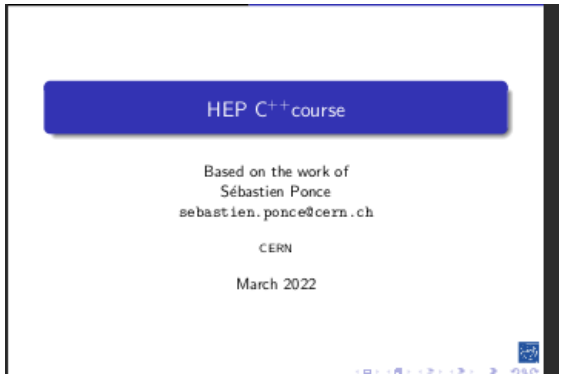
Data Gathering: A Tool Comparison

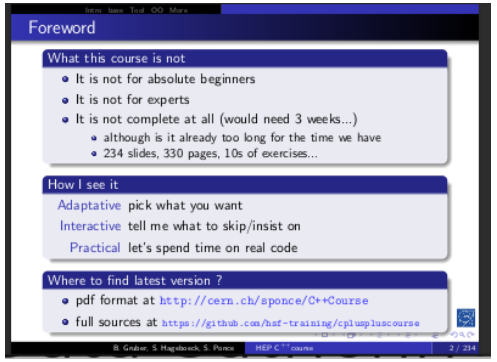
CDS
Papers
& Notes

Slides more difficult to process
and largest data in volume









Faster

Marker vs Nougat
(PDF scraping)

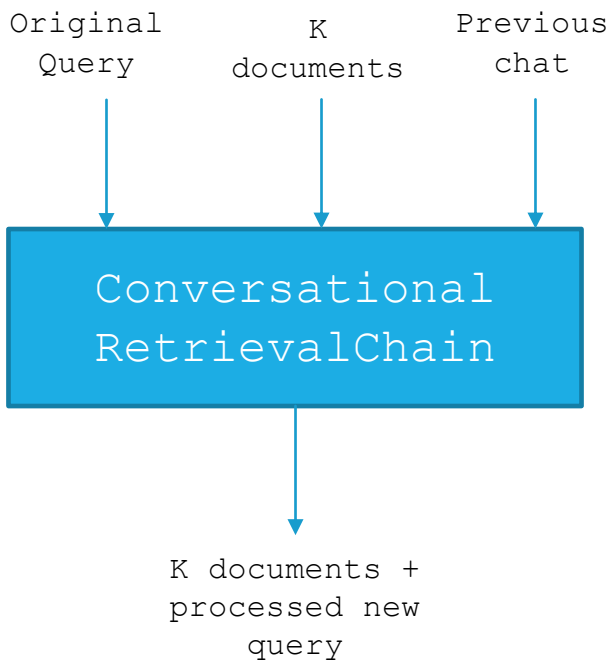
Processing: Chunking and Retrieval

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

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



Summary of Gathering and Processing

Summary includes...

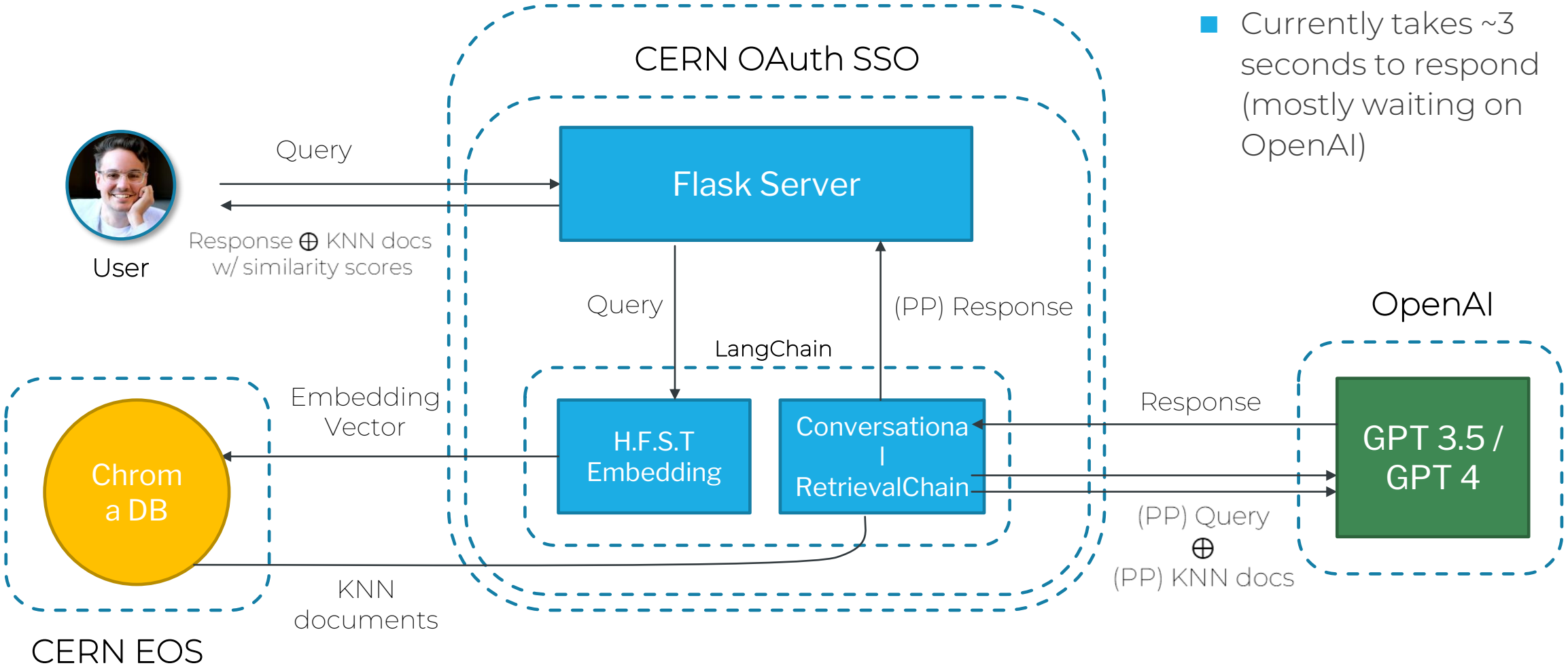
- Diagram of all the possible ATLAS datasets and how many we have in the Database(DB)
- Chunked & Embedded Datasets are ready for or have been added to DB
- Stage the progress of each
- Over 37% of textualizable ATLAS datasets (minimum)
- Hardware upgrades were made to scrape in larger volume

Task \ **DB**	Twiki	ATLAS Software Docs	E-group Archive	Indico Meetings	PDF Plots	Mattermost	Jira	ATLAS Codebases	Group level Docs	CDS Papers & Notes
Scrape	2k+	500+	10k+	~1k+	~1k+					~5k+
Convert										
Chunk & Embed										

Not yet started
In Progress
Complete

Current INFRASTRUCTURE

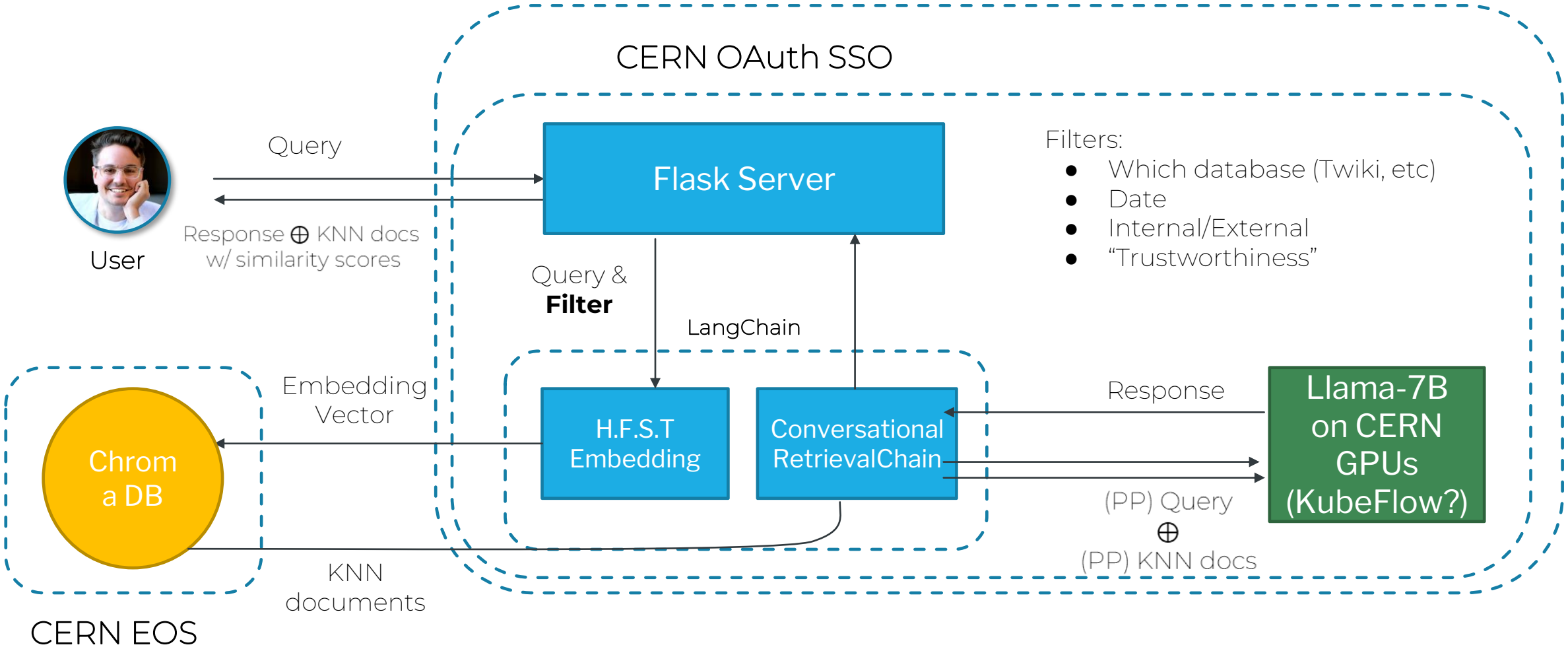
(PP) = (Possibly Processed)



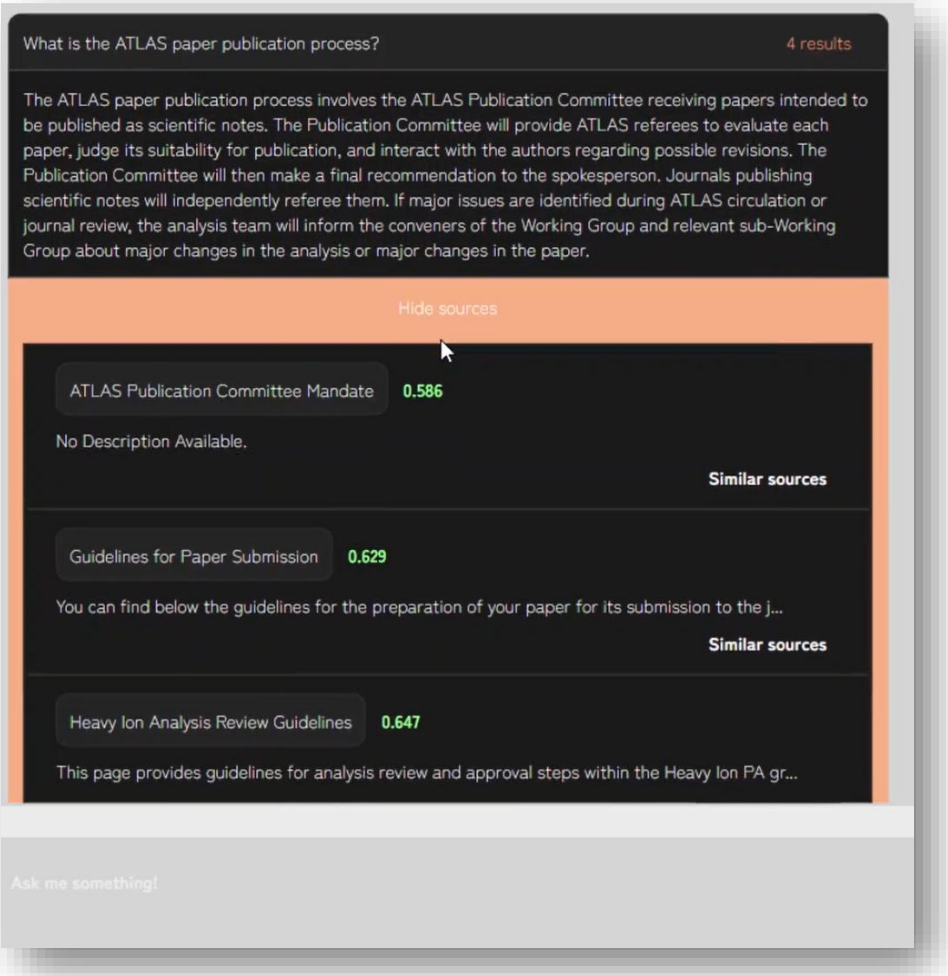
- Currently takes ~3 seconds to respond (mostly waiting on OpenAI)

Planned Infrastructure

(PP) = (Possibly Processed)

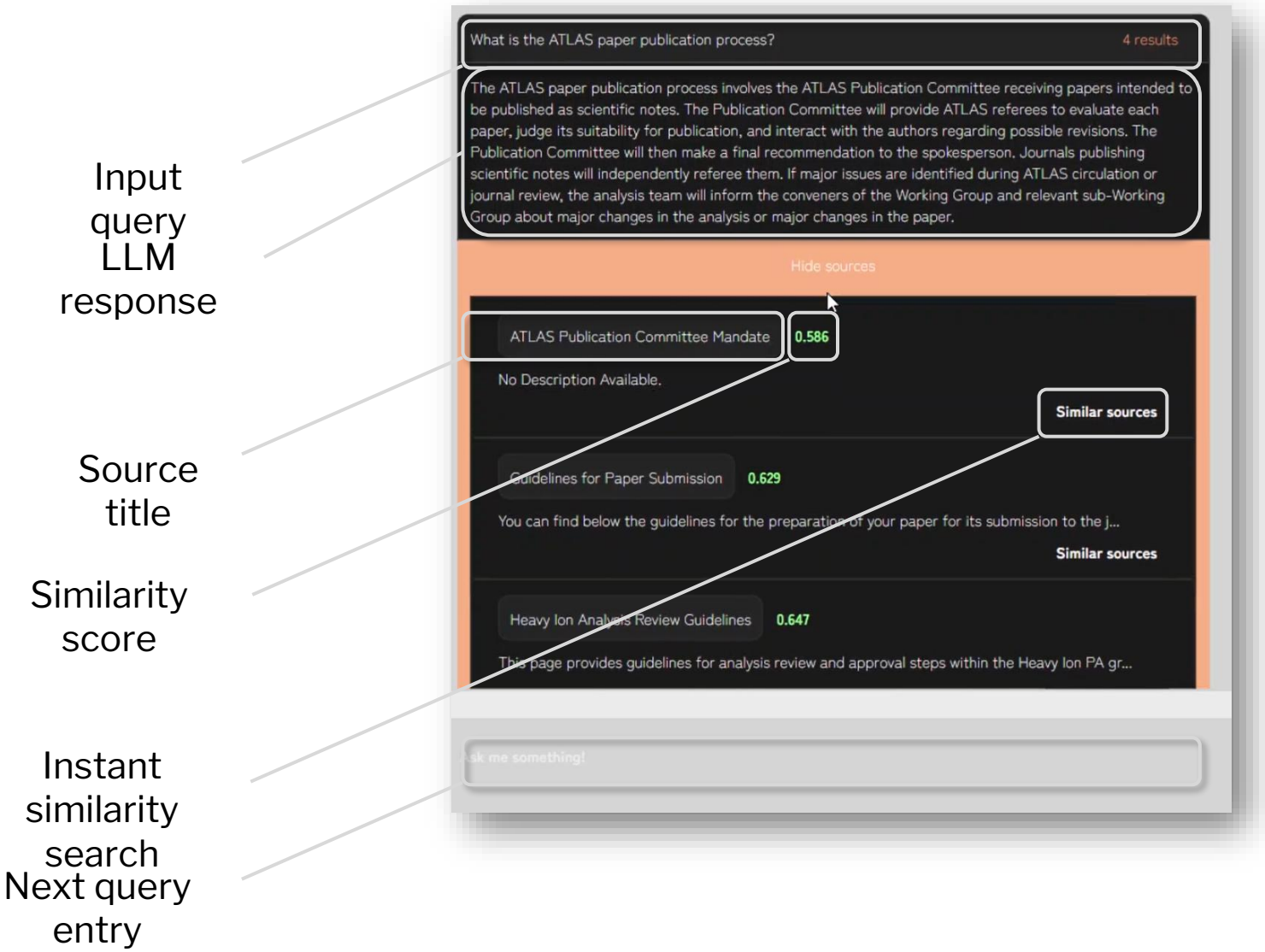


Chat Interface



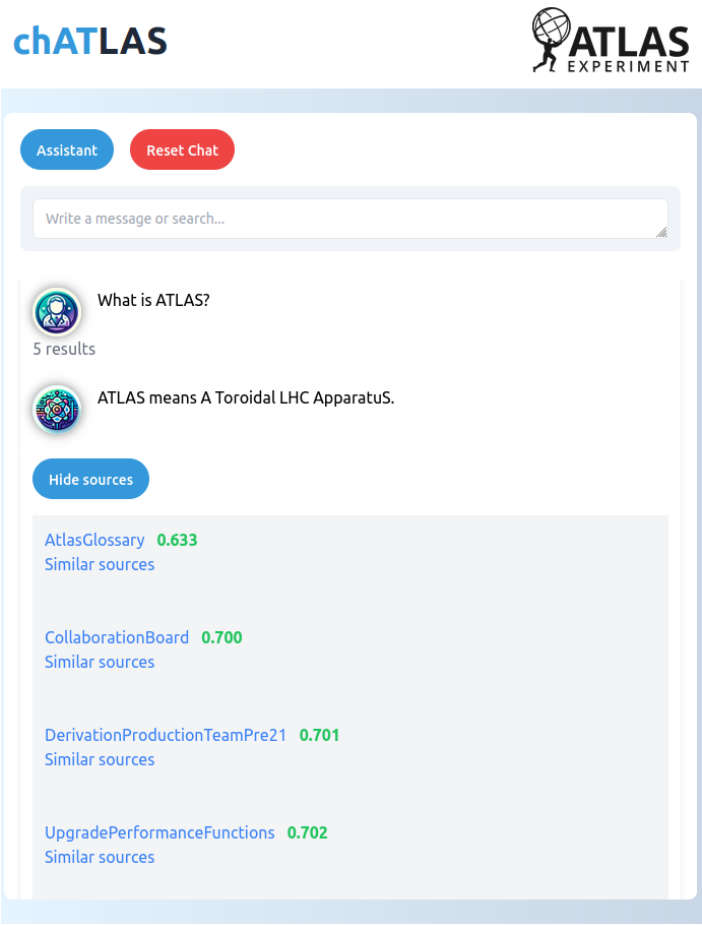
Chat Interface

- 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



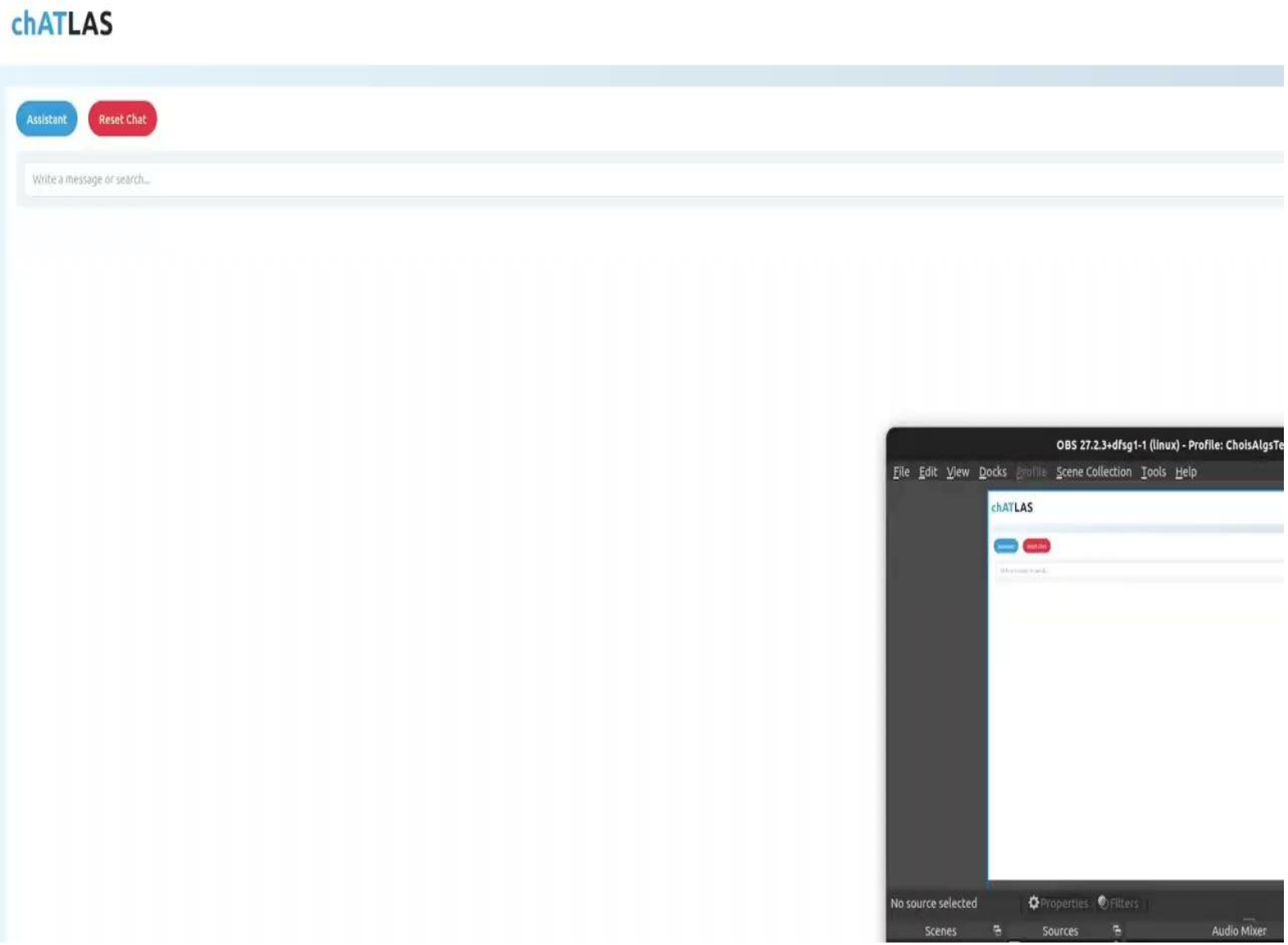
Chat Interface

- 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

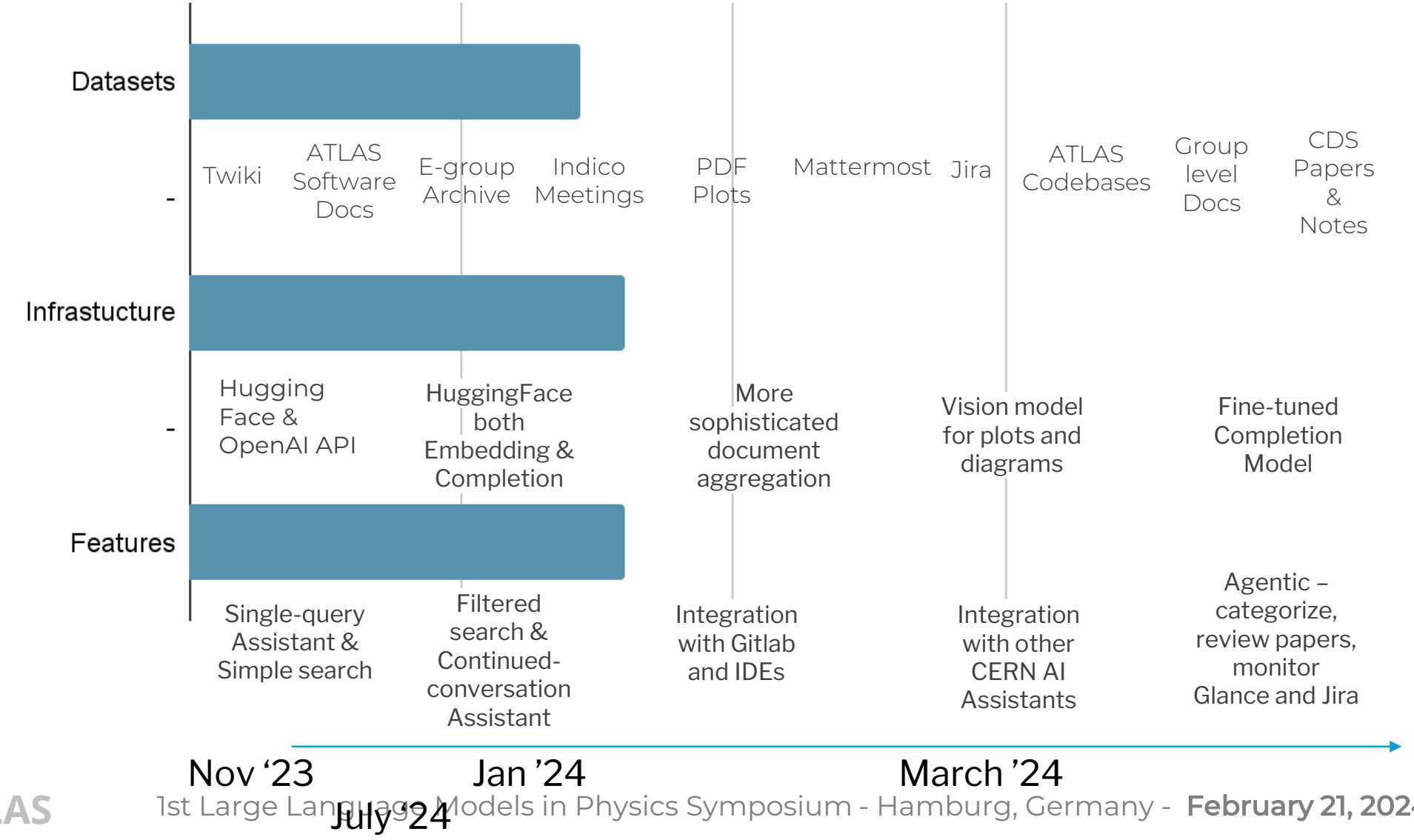


This is a chatbot application developed for ATLAS and CERN. It operates in both assistant and search modes across a variety of ATLAS databases (currently Twiki).

Demo


chATLAS
Assistant
Reset Chat
Write a message or search...
OBS 27.2.3+dfsg1-1 (linux) - Profile: ChoisAlgsTe
File Edit View Docks Profile Scene Collection Tools Help
chATLAS
No source selected
Properties
Filters
Scenes
Sources
Audio Mixer

Roadmap (Milestone Tracking)



Recent Updates and Ongoing Work

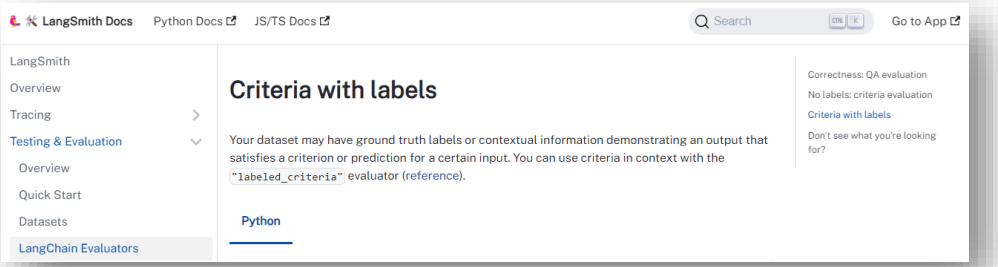
- Studies of the use and performance of different models as well as *response optimization* via *prompt engineering* and *similarity search techniques*
- Indico scraping shown previously with Nougat and Marker
- Video caption search and embedding
- Fine-tuning model research and experiments on either CDS or Twiki
- App development, deployment, documentation, etc
- Studies on how to include the ATLAS Glossary, and other “Dictionary” or “Reference” type documents - first-layer database search, fine-tuning, or even heuristic hand-engineering (“When is the NSW being upgraded?” - search in Glossary for any words present - place them in parentheses = “When is the NSW (New Small Wheel) being upgraded?”)

Challenges and Suggestions

- 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 ~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?

OPEN QUESTIONS

- How to avoid hallucinations? (Integrate latest research?)
- 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? (Metric Development)
- What is the best dataset to gather for fine-tuning?
- How to anonymize email threads and discussion forums?



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!

Presentation Summary

- Our goal is to create a reliable AI assistant across all ATLAS content
- Solved(In Progress) - ATLAS has significant data and presents a logistical challenge that we have largely overcome
- Complete - Implemented good semantic (vector embedded) search
- Complete - Prototype of AI Assistant
- Roadmap - Achieving our goal by Iterative Development
- Feedback - Recent Updates, Challenges, Suggestions, and Open Questions

END