chATLAS

An AI Assistant for the ATLAS Collaboration

Jeremy Couthures[1], Daniele Del Santo[2], Gabriel Facini[3], Joe George[3], Runze Li[4], Daniel Murnane[5,6], Cary Randazzo[7]

IML Meeting, 9th April 2024

(1) LAPP
(2) UNIVERSITÄT BERN
(3) UCL
(4) Yale
(5) BERKELEY LAB
(6) UNIVERSITY OF COPENHAGEN
(7) LOUISIANA TECH UNIVERSITY

Introduction

Large language models (LLMs), on the back of the transformer architecture revolution, will continue to permeate society. Game changer.

Vector Databases are an integral feature of the ecosystem

- Store mathematical representations of data in a high-dimensional space
- Can be used to limit hallucinations
- Basis of “Ask the docs” or RAG chat bots

ATLAS documents: disperse & difficult to navigate:

- Which TWiki has the information that will save me a month of work?
- Which papers have a Z control region and correct the simulation?
- Which e-group can I check for answers to my question?

Motivation: Can we use this modern toolkit to accelerate research?
A few of us are exploring pragmatic pathways to leverage open source tools with realistic expectations in our spare time.

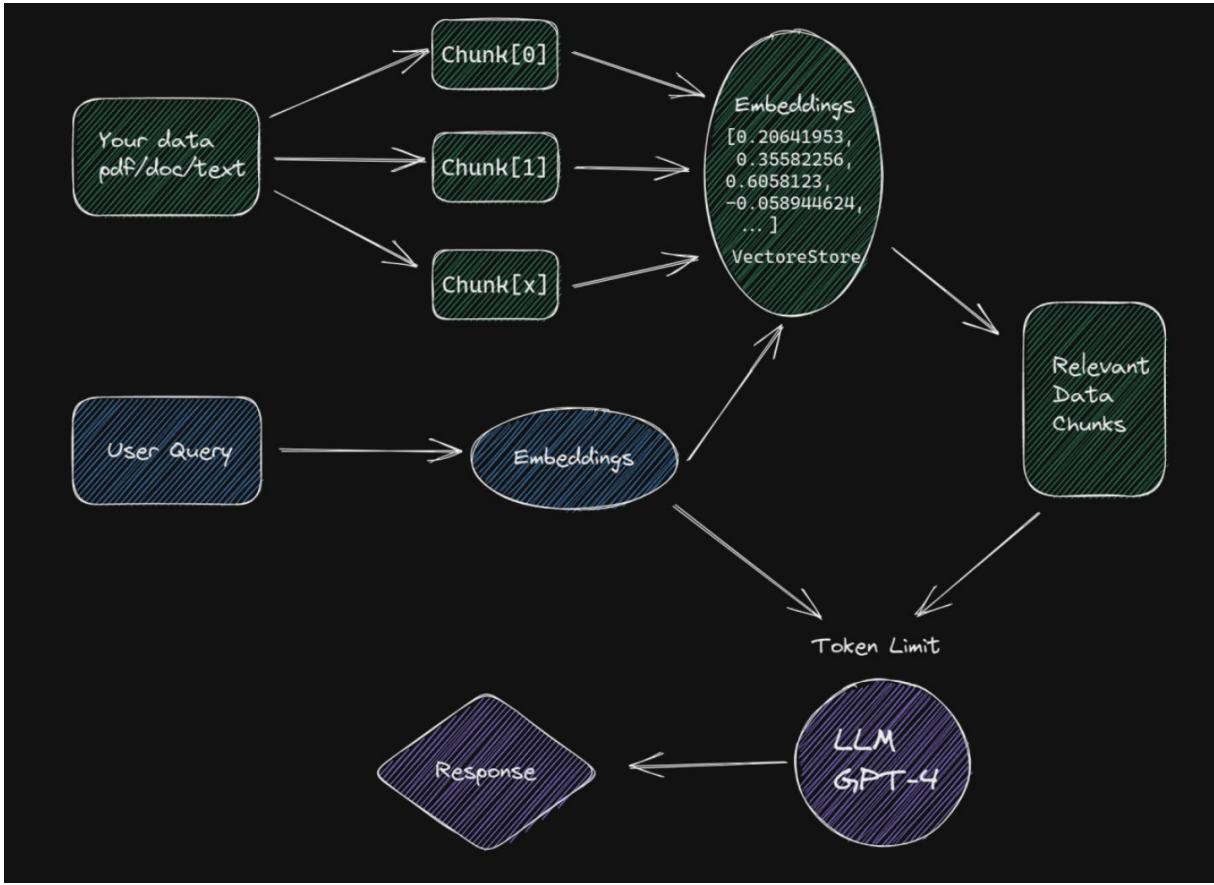
Retrieval-Augmented Generation (RAG) model

Goal: Ask questions to an LLM about private information *without retraining*

Setup: Embeddings: Convert sentences into vectors in a multidimensional space

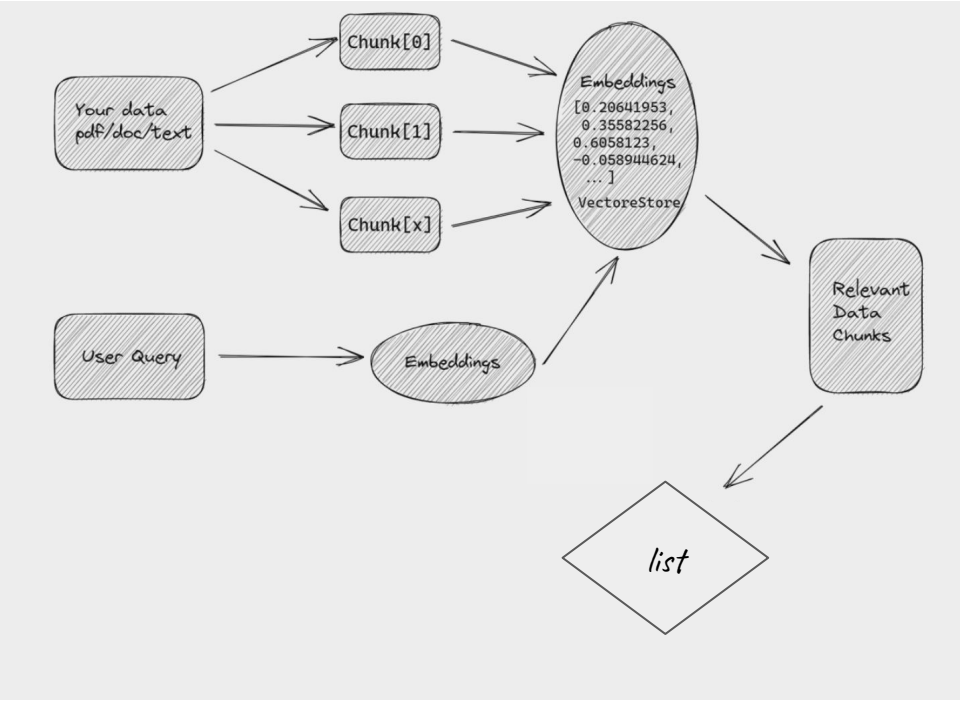
Usage:

1. Query vector space
2. Send query+docs to LLM
3. LLM formulates response

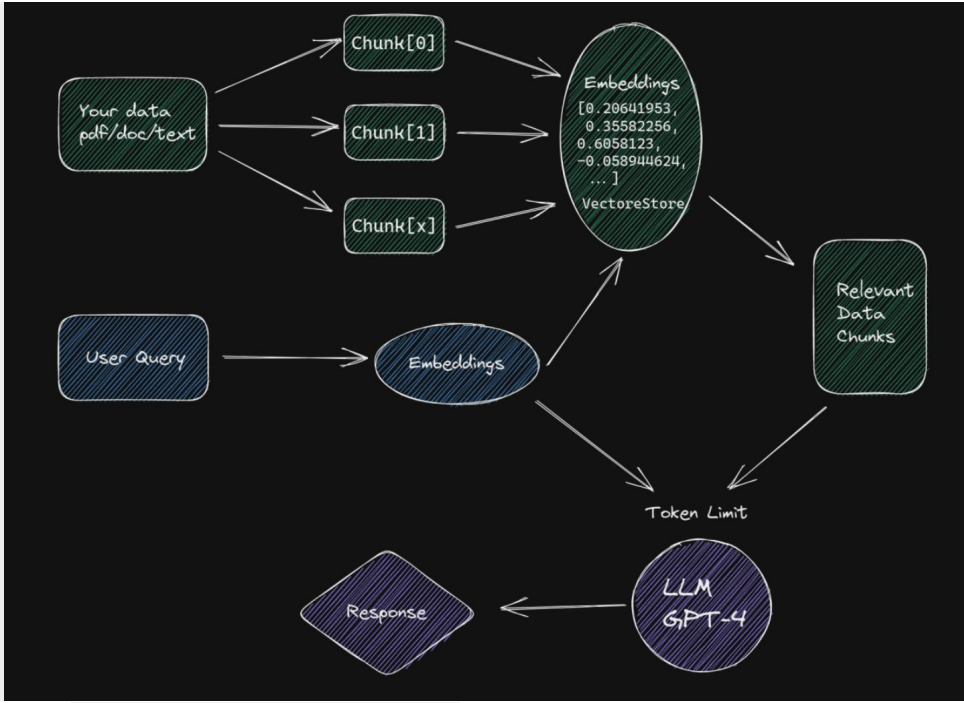


chATLAS: RAG using various internal ATLAS sources

Search mode (not a RAG)



Assistant (RAG)



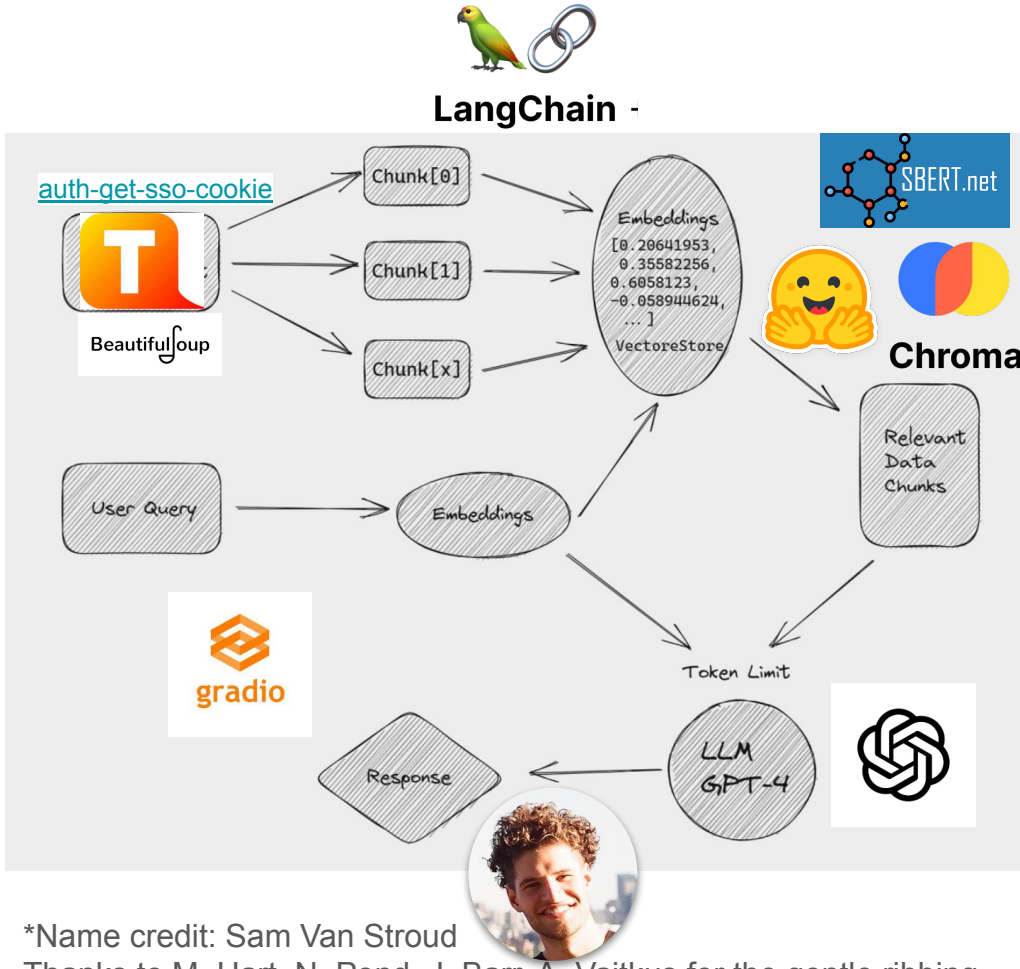
Simple to start

June 2023, ATLAS ML Forum meeting

- ATLAS-GPT: Daniel Murnane
- chATLAS*: Gabriel Facini
- Consolidated behind the (potentially) free option

DM

GF

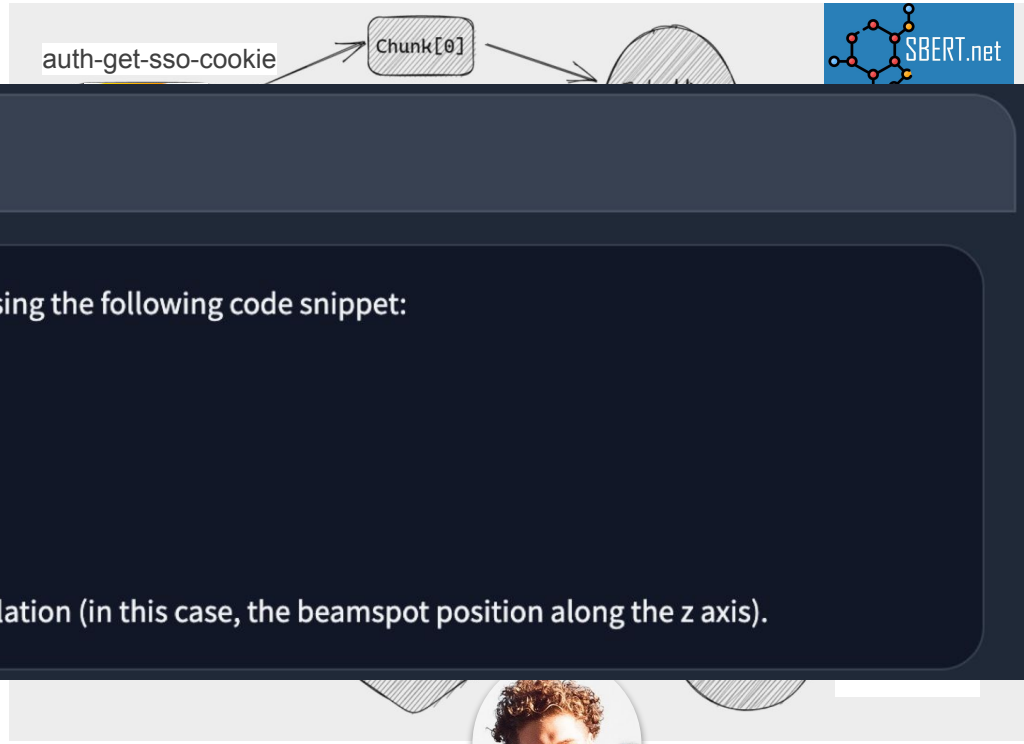


*Name credit: Sam Van Stroud

Thanks to M. Hart, N. Pond, J. Barr, A. Vaitkus for the gentle ribbing

Simple to start

LangChain -



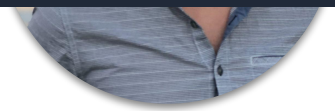
How do I calculate the z0 of a track with respect to a vertex?

You can calculate the z0 of a track with respect to a vertex by using the following code snippet:

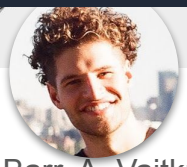
```
xAOD::TrackParticle trk;
xAOD::Vertex vtx;
float delta_z0 = fabs(trk.z0() + trk.vz() - vtx->z());
```

where trk.vz() represents the point of reference for the z0 calculation (in this case, the beamspot position along the z axis).

DM



GF



*Name credit: Sam Van Stroud

Thanks to M. Hart, N. Pond, J. Barr, A. Vaitkus for the gentle ribbing

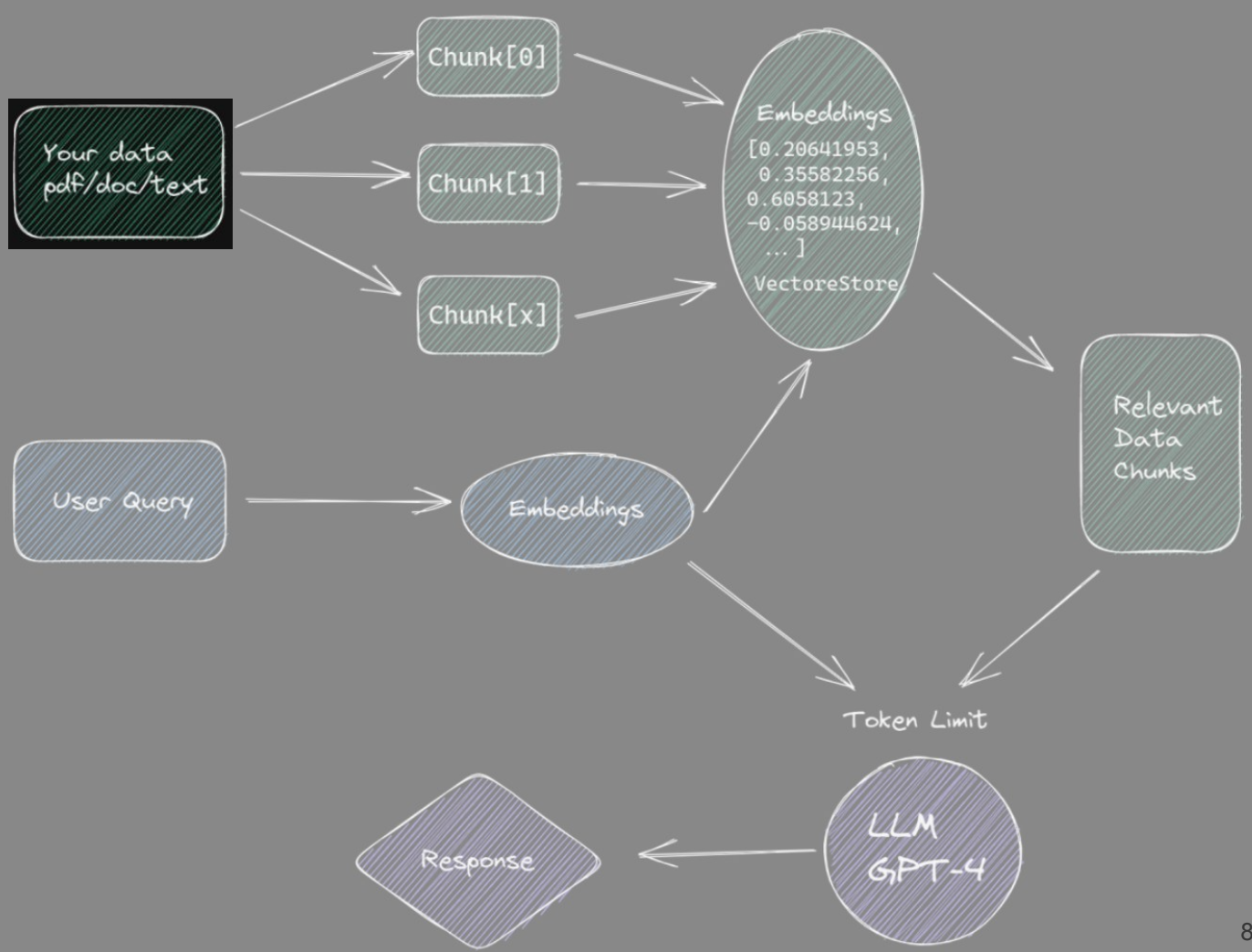
More Info to User

Moved UI to Flask to add more information:

- expose search results…
- …with similarity score
- Can follow source with similarity search.

JG

1. Scraping & Document Collection



Expanding footprint: + more TWikis +CDS

RL

Significant expansion of collected dataset:

- Expand initial scrape of TWikis. Expand from 1.5k to ~22k
- Add 1.2k ATLAS published papers from CDS
- META's Neural Optical Understanding for Academic Documents (NOUGAT) converts PDFs to markdown preserving math symbols and formulae
  - **DB with CDS docs in alpha testing**

Room to grow. Some low hanging fruit:

- TWikis: Do not include software docs (yet)
- CDS: +66k records, 36k internal. Room to (easily) grow

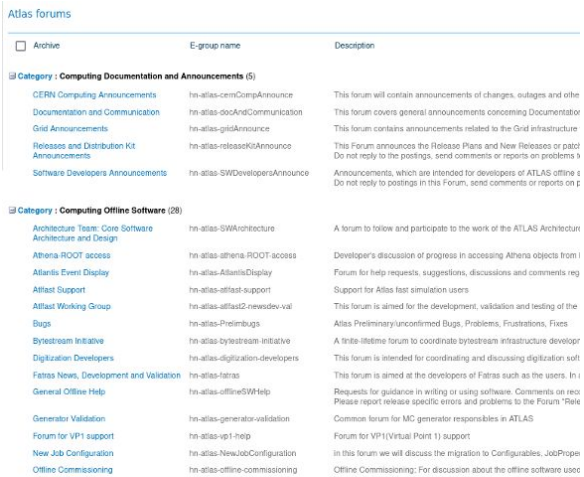
Latest scrapers are here (2FA blocked)

Expanding footprint:+e-groups

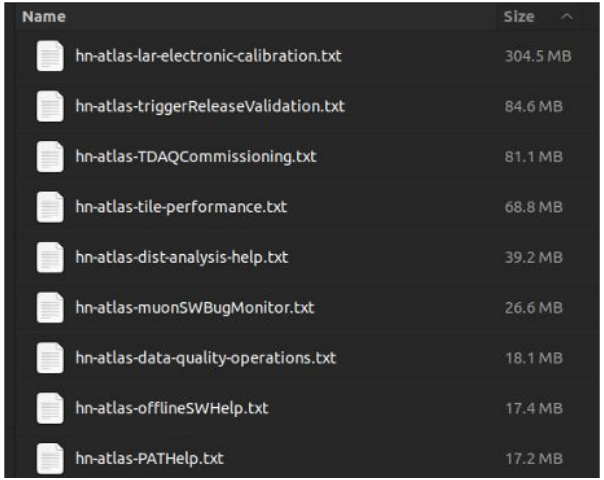
CR

Using **Selenium** & **BeautifulSoup**, cracked e-groups

- 8912 topics x (1-50+ messages) for largest egroup
- ~ tens of thousands of messages to date
- **Next step:** Embed and include!



Code
(in alpha)



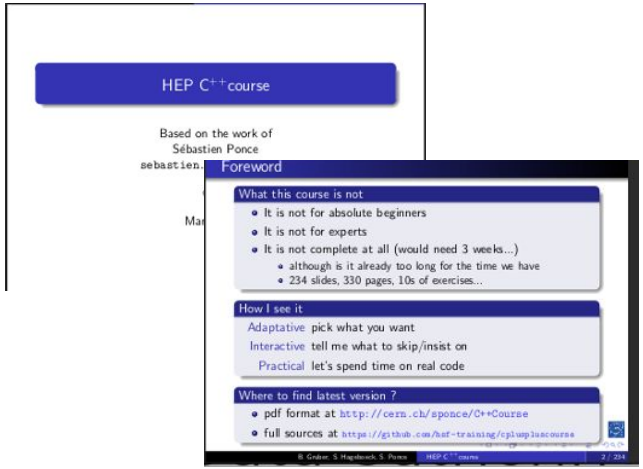
BeautifulSoup

Expanding footprint:+indico

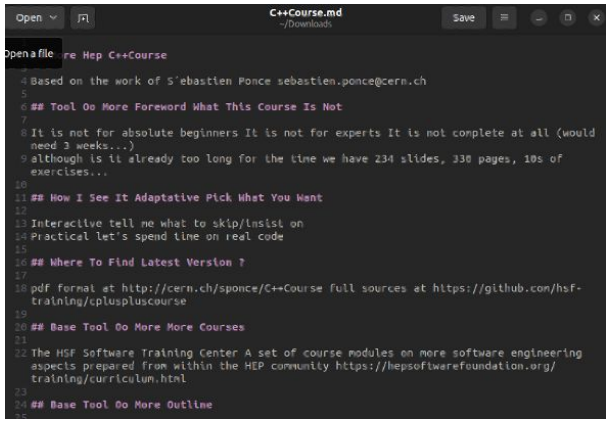
DDS

Using NOUGAT and marker, pipeline established
As less structured, difficult to generally include
Initial targets are well prepared slides + audio recordings:

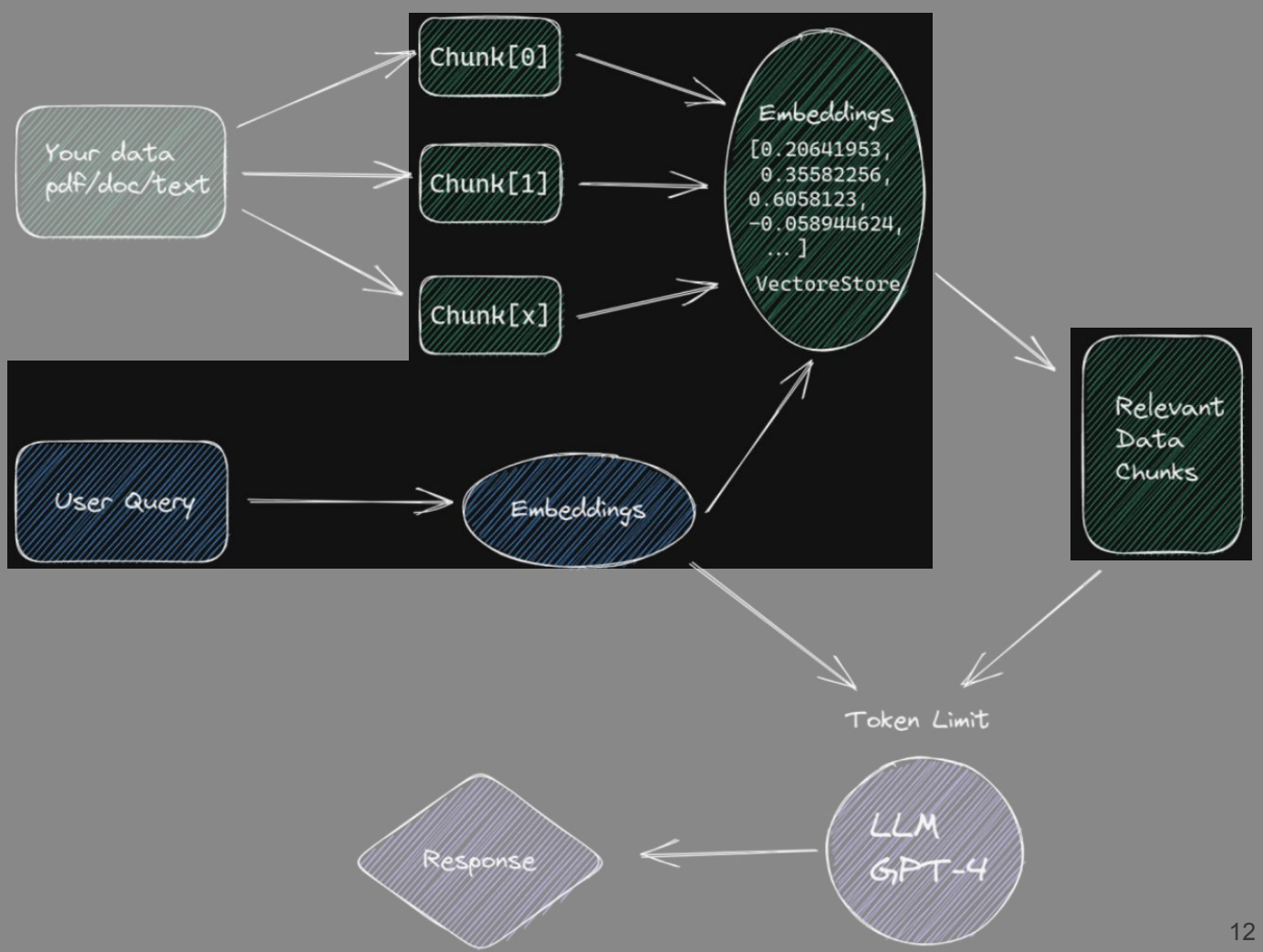
- i.e: ATLAS Weekly, Tutorials, ECSB events i.e. Induction Day
- CERN IT has custom audio to text model, specific for HEP lectures/talks



Code
(In alpha)



2. Embedding & Vector Store



Embeddings

GF

Model dependent. Using: [sentence-transformers/all-MiniLM-L6-v2]

- Small, runs on a local CPU
- Multiple “chunks” from one document
- Plan: explore larger or custom models

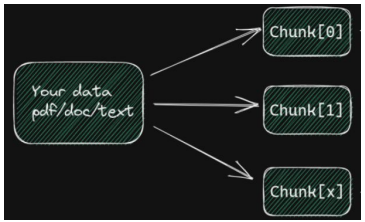
Search

Chroma

RL introduced a child-parent structure:

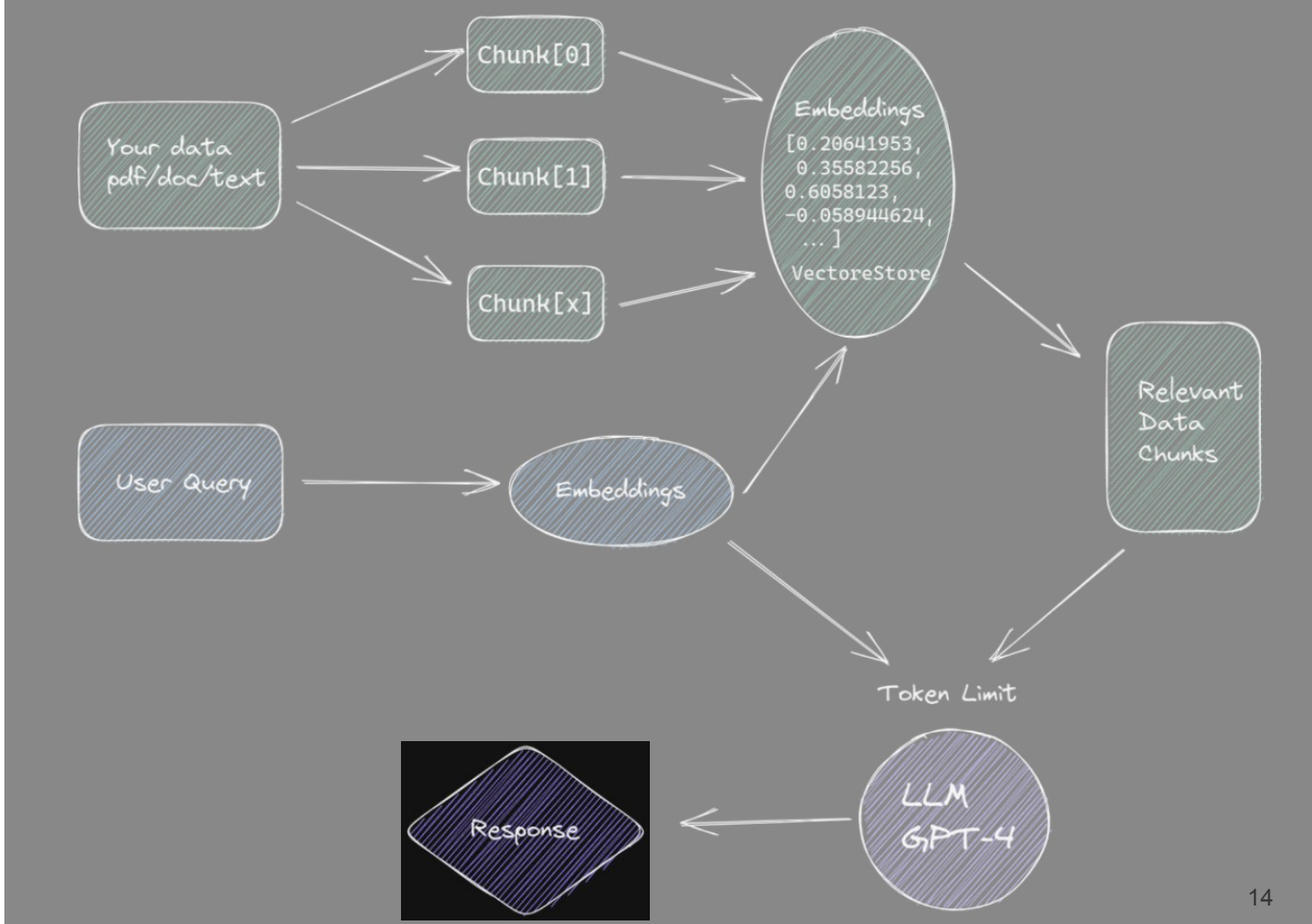
- **Parent** is full doc. **Children**: Chunks of ~200 tokens
- Search children, parent to LLM

Improved results on basic tests



RL

3. Evaluation



How good is it?

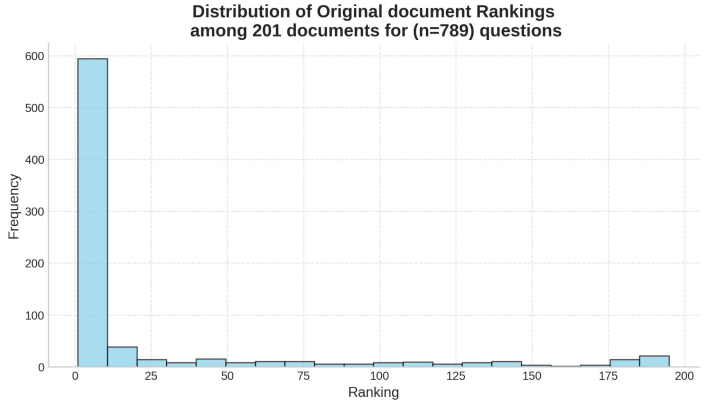
JC

Quality control, test performance, test evolutions
JC got the work started with generating test data:

- For each TWiki, ask an LLM to list X questions that can be answered by the TWiki & response

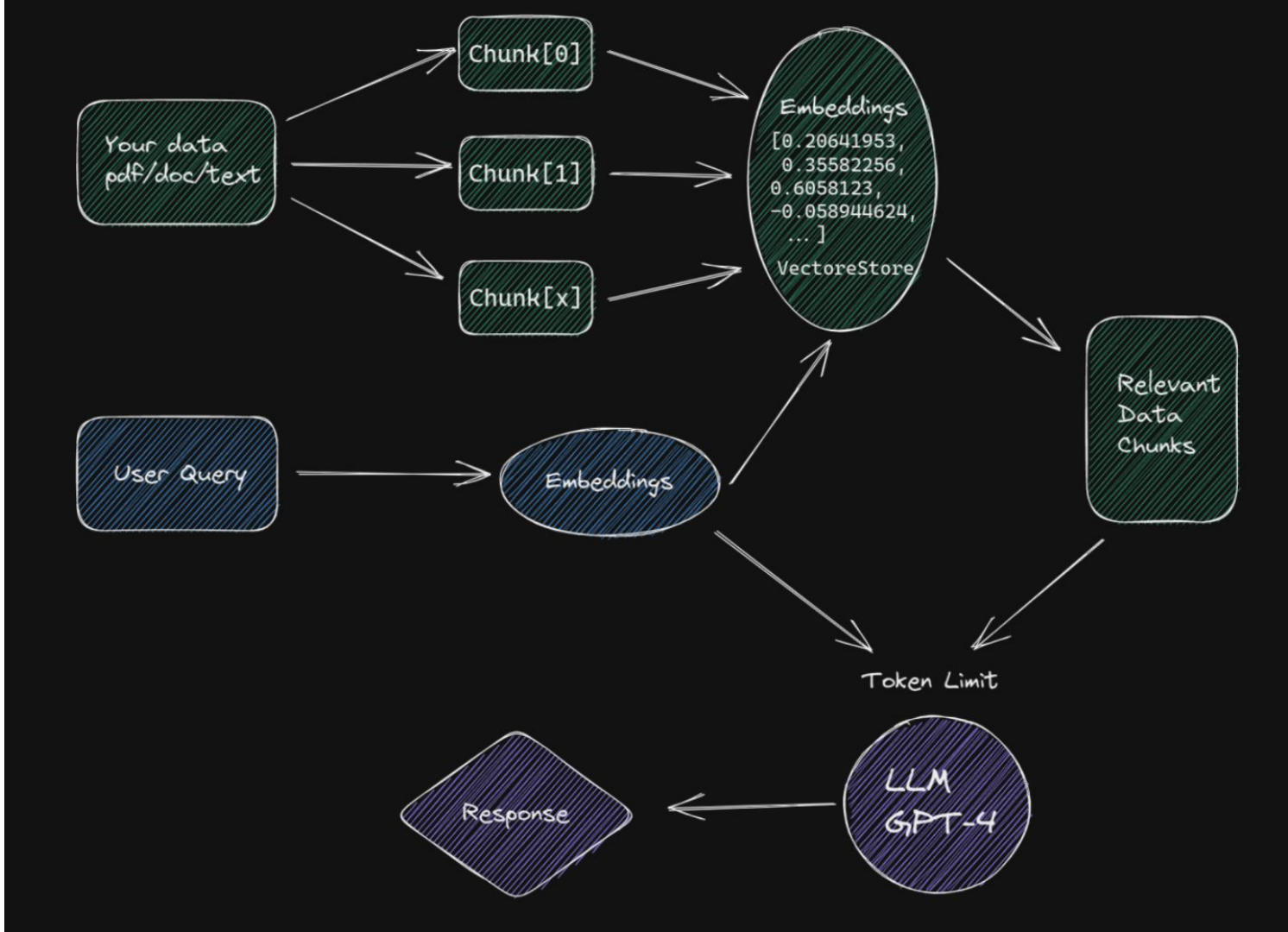
Invert for testing:

- **Search**: Ask question and get the rank of the TWiki among the documents (see plot)
  - **With current setup, correct doc returned ⅔ of time**
- **Assistant**: Ask question and see if answer matches (cosine sim) the “known” answer

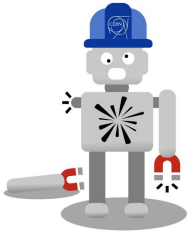


Small test: Ability to find correct TWiki using 789 questions from 201 TWikis

4. UI & Serving in Production



chATLAS in Production



DM tackled CERN's Platform-as-a-Service (PaaS)

- Driven from production git repo
- Launched in Nov. ‘23, stability improved, still in early testing
- Working with IT Dept’s **IT-GPT** project, will report further details in tomorrow’s meeting
- Currently runs on entire ATLAS twiki. CDS and egroups coming ASAP
- (For power users) Can configure search and assistant (see backup for details)

DM

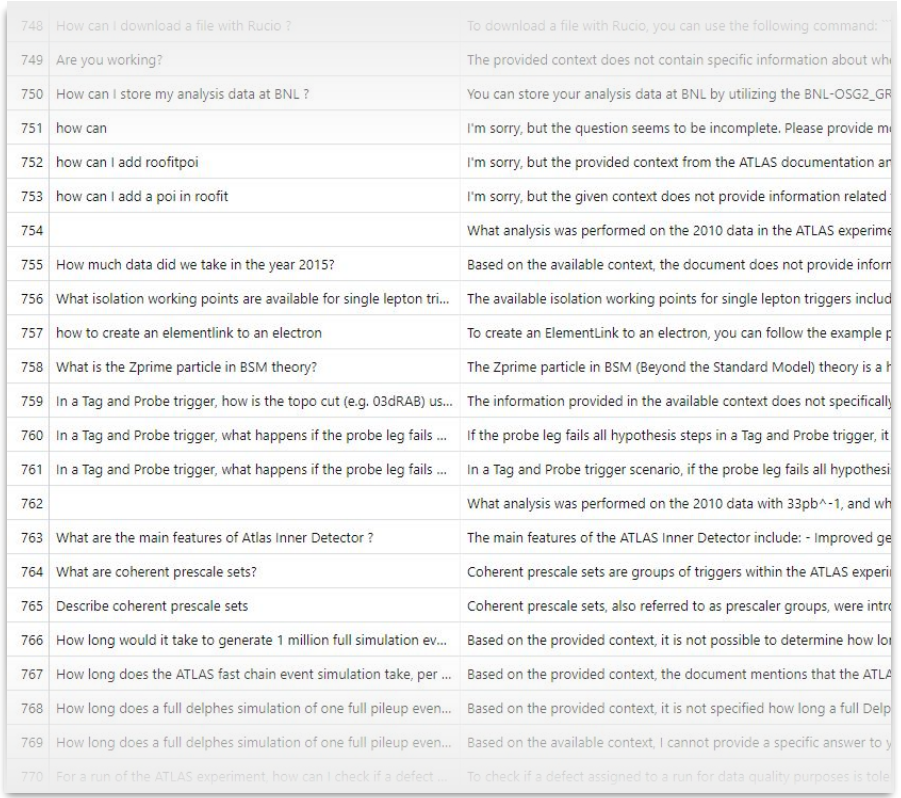
okd
PaaS Docs

Available for testing now!

Please try it out and give feedback atlas-comp-ml-chatlas-developers@cern.ch

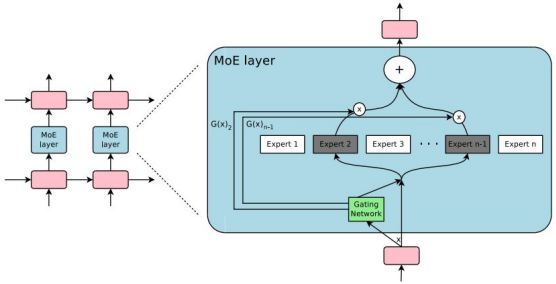
chATLAS in Production

- Since public beta launch ~2 weeks ago, there have been almost 1000 queries to the assistant
- Around 40% of these appear (from human inspection) to be decent answers
- The remainders are typically a refusal to answer due to lack of context, meaning that our retrieval system needs to be improved
- Already see improved context using Maximal Marginal Relevance retrieval
- Exploring:
  - More sophisticated automatic prompt generation
  - Passing query through ATLAS glossary
  - More physics-friendly embedding models



748	How can I download a file with Rucio ?	To download a file with Rucio, you can use the following command: ```
749	Are you working?	The provided context does not contain specific information about wh
750	How can I store my analysis data at BNL ?	You can store your analysis data at BNL by utilizing the BNL-OSG2_GR
751	how can	I'm sorry, but the question seems to be incomplete. Please provide m
752	how can I add roofitpoi	I'm sorry, but the provided context from the ATLAS documentation an
753	how can I add a poi in roofit	I'm sorry, but the given context does not provide information related
754		What analysis was performed on the 2010 data in the ATLAS experime
755	How much data did we take in the year 2015?	Based on the available context, the document does not provide inform
756	What isolation working points are available for single lepton tri...	The available isolation working points for single lepton triggers includ
757	how to create an elementlink to an electron	To create an ElementLink to an electron, you can follow the example p
758	What is the Zprime particle in BSM theory?	The Zprime particle in BSM (Beyond the Standard Model) theory is a h
759	In a Tag and Probe trigger, how is the topo cut (e.g. 03dRAB) us...	The information provided in the available context does not specifically
760	In a Tag and Probe trigger, what happens if the probe leg fails ...	If the probe leg fails all hypothesis steps in a Tag and Probe trigger, it
761	In a Tag and Probe trigger, what happens if the probe leg fails ...	In a Tag and Probe trigger scenario, if the probe leg fails all hypothesi
762		What analysis was performed on the 2010 data with 33pb^-1, and wh
763	What are the main features of Atlas Inner Detector ?	The main features of the ATLAS Inner Detector include: - Improved ge
764	What are coherent prescale sets?	Coherent prescale sets are groups of triggers within the ATLAS experi
765	Describe coherent prescale sets	Coherent prescale sets, also referred to as prescaler groups, were intro
766	How long would it take to generate 1 million full simulation ev...	Based on the provided context, it is not possible to determine how lo
767	How long does the ATLAS fast chain event simulation take, per ...	Based on the provided context, the document mentions that the ATLA
768	How long does a full delphes simulation of one full pileup even...	Based on the provided context, it is not specified how long a full Delp
769	How long does a full delphes simulation of one full pileup even...	Based on the available context, I cannot provide a specific answer to y
770	For a run of the ATLAS experiment, how can I check if a defect ...	To check if a defect assigned to a run for data quality purposes is tole

Near term plans/goals



MISTRAL AI_

1. Incorporate more sources:
   a. CDS papers (in alpha), confs, pubs, INTs
   b. e-groups (scraped, to be embedded)
   c. Indico (some sources in hand)
2. Daily updates i.e. from TWiki edits
3. Find my plot - find plots more easily by storing information in DB
   Running - need to serve via API
4. Fine tune of open source LLM a la AstroLLaMA - low stats show promise
5. Serving open source LLM via small API running at university
6. Improve quality through advanced RAG techniques
   a. Or mixture of experts like ChatGPT
7. UI improvements: thumbs up/down etc

Invitation to Contribute: The chATLAS Challenge

90% of students will not end in HEP. We are funded as a training ground for skills valuable in the economy. A skill gap exists. We offer an opportunity to developing familiarity and skills with some of the tools around LLMs.

Goals:

- Establish leaderboard based on performance metrics (see previous slides)
- Regularly adopt models & methods which perform better
- Allow users to develop and deploy new features
- Will announce best-performing models at ATLAS ML Workshop in May

Resources:

- development repo
- example notebooks
- production repo - *contains instructions to run locally or on lxplus if PaaS is unresponsive*
- dataset locations

Conclusions

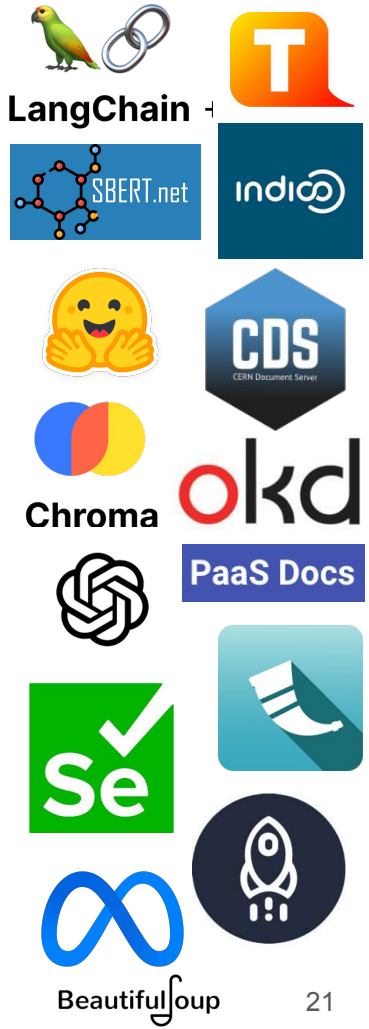
Finding information inside ATLAS is hard. Made efficient, collectively we can achieve more and potentially discover new physics sooner.

- The LLM revolution offers us tools that can facilitate knowledge sharing - i.e. Retrieval-Augmented Generation (RAG) models

chATLAS project established in **spare time** of a few collaborators

- Barely off the ground, and are pushing towards a stable state
- PaaS platform stability is the biggest issue on hand
- Announcements e-group: atlas-comp-ml-chatlas-information

Open invite to contribute. No neg bombs, develop instead



BACKUP

Privacy Concerns

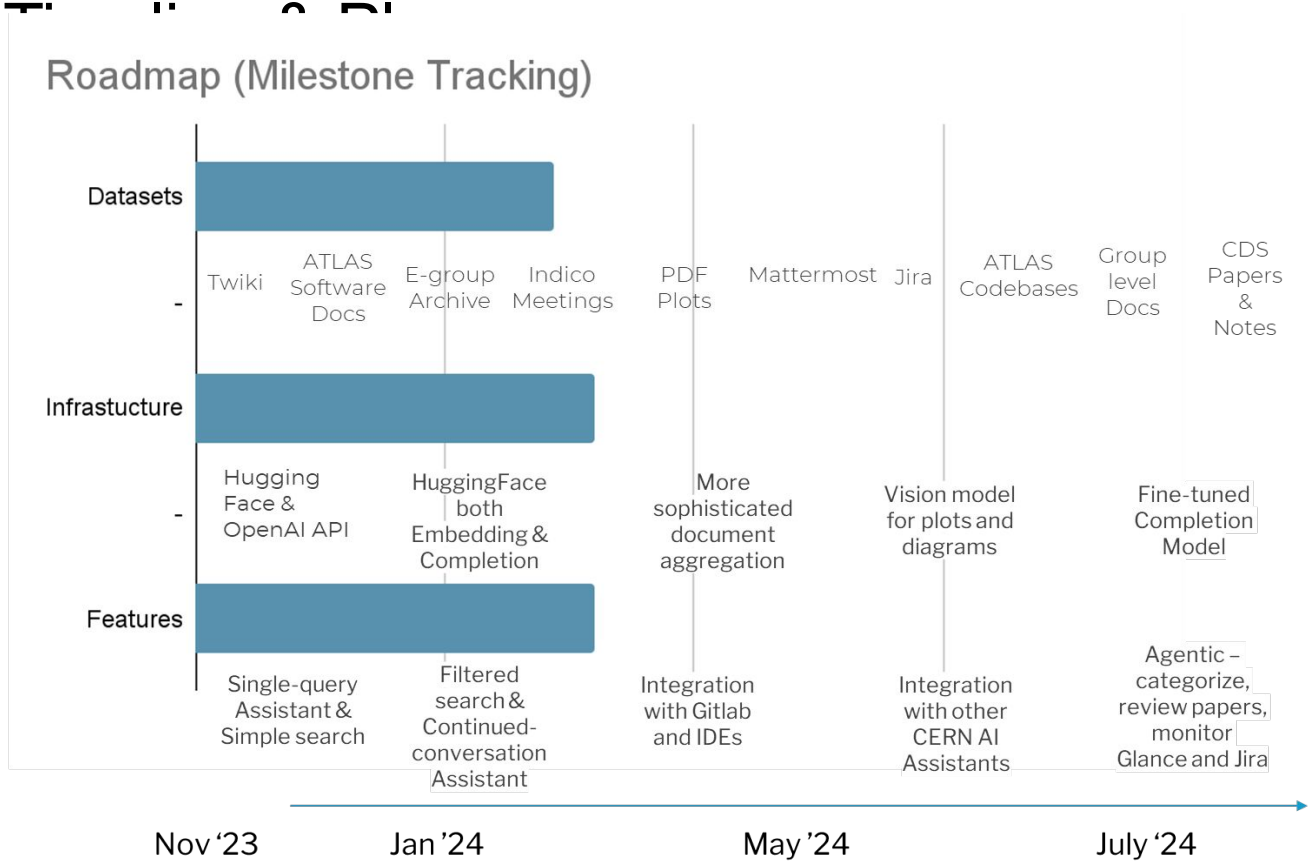
- Model used to embed run on local CPU [no issue]
- Using OpenAI API transmits data to OpenAI. Data Policy
  - As of March 1, 2023, data sent to the OpenAI API will not be used to train or improve OpenAI models (unless you explicitly opt in)
  - To help identify abuse, API data may be retained for up to 30 days, after which it will be deleted (unless otherwise required by law).
- Audio recordings: target structured presentations, not unstructured discussion
  - Ideal: Words someone uses to explain a tutorial.
  - Q&A in e-group is captured in searchable archive. Can replicate within chATLAS DBs
  - Q&A in meeting is not captured in searchable archive. To respect more informal forum, avoid this type of information within chATLAS DBs
- Query logging: We log questions asked without identifiable user information (i.e. usernames are hashed to preserve anonymity)

Specific Timeline & Plans

DB / Task	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

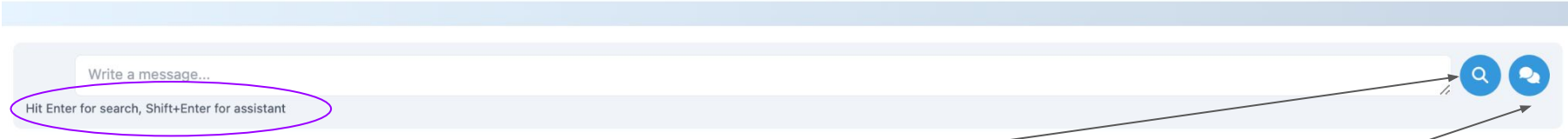
Specific T[illegible]


Roadmap (Milestone Tracking)
Datasets
Twiki
ATLAS Software Docs
E-group Archive
Indico Meetings
PDF Plots
Mattermost
Jira
ATLAS Codebases
Group level Docs
CDS Papers & Notes
Infrastucture
Hugging Face & OpenAI API
HuggingFace both Embedding & Completion
More sophisticated document aggregation
Vision model for plots and diagrams
Fine-tuned Completion Model
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
Nov '23
Jan '24
May '24
July '24

Overview of App

Configure (see next slide for options)

Reset history



Enter = search mode. Lists Twikis (with links) in order of search strategy (see configurations)

Shift+Enter = assistant mode. See response from LLM (time delay to talk to OpenAI) with option to see search results

Power Users: Configuring the Production Application

To config, click gear