

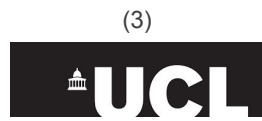
chATLAS

An AI Assistant for the
ATLAS Collaboration



Jeremy Couthures¹, Daniele Del Santo², Gabriel Facini³, Joe George³,
Runze Li⁴, Daniel Murnane^{5,6}, Cary Randazzo⁷

IML Meeting, 9th April 2024



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.



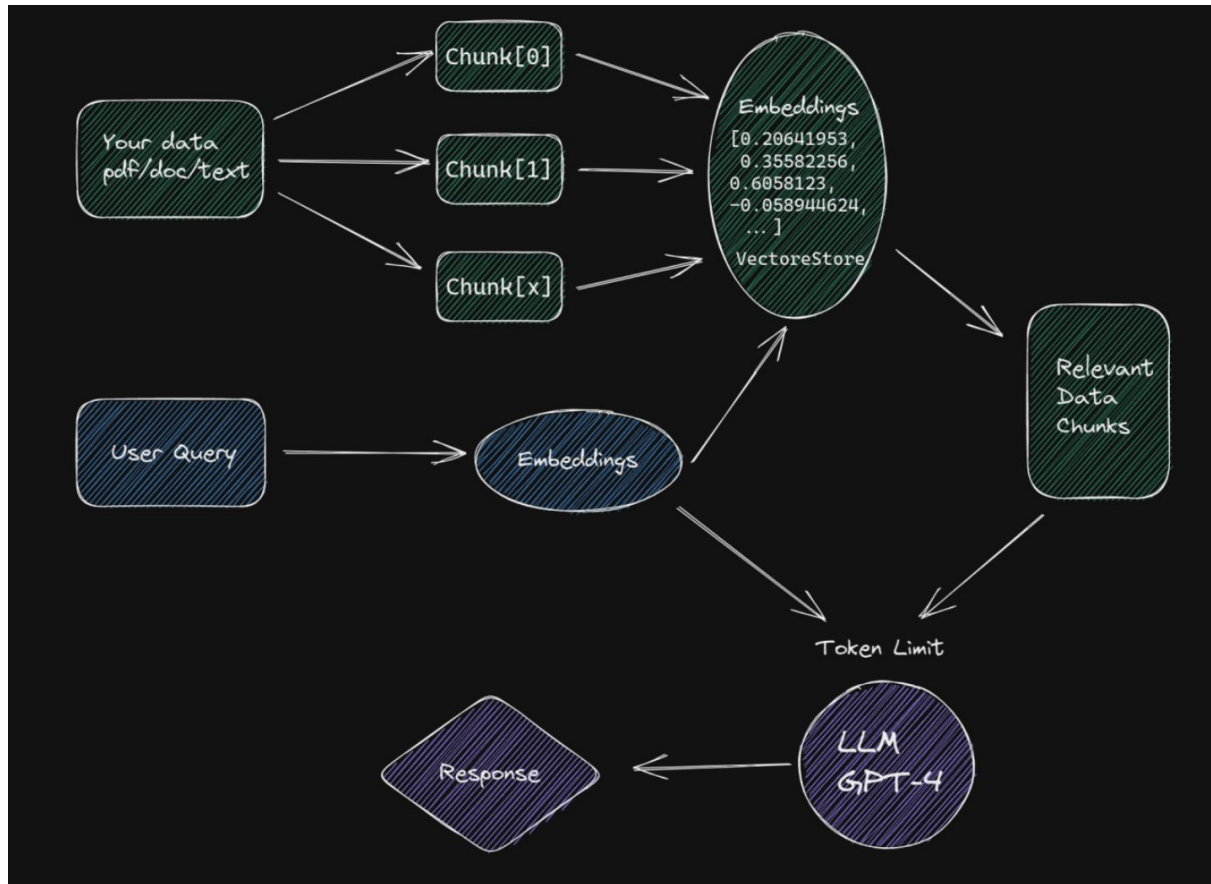
Retrieval-Augmented Generation (RAG) model

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

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

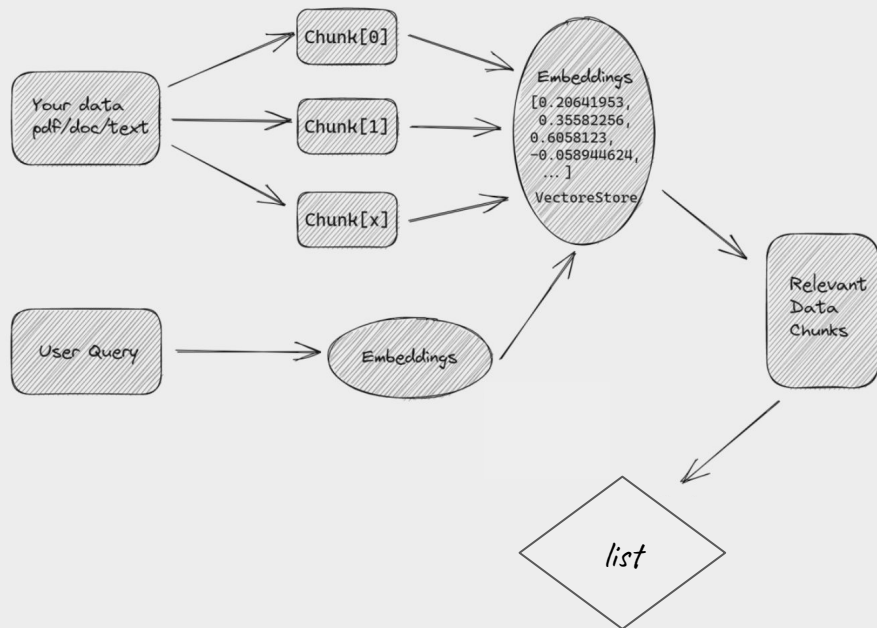
Usage:

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

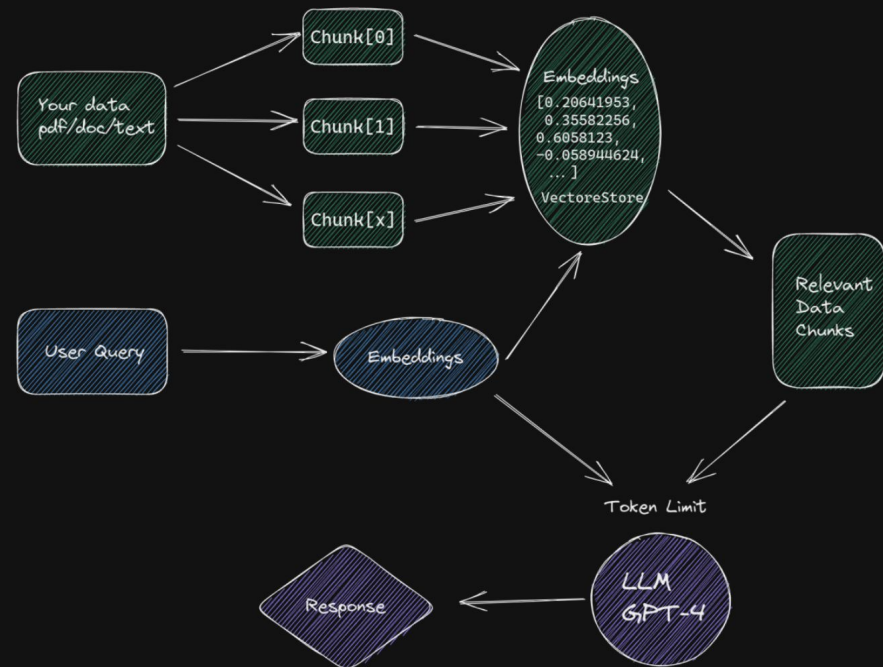


chATLAS: RAG using various internal ATLAS sources

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

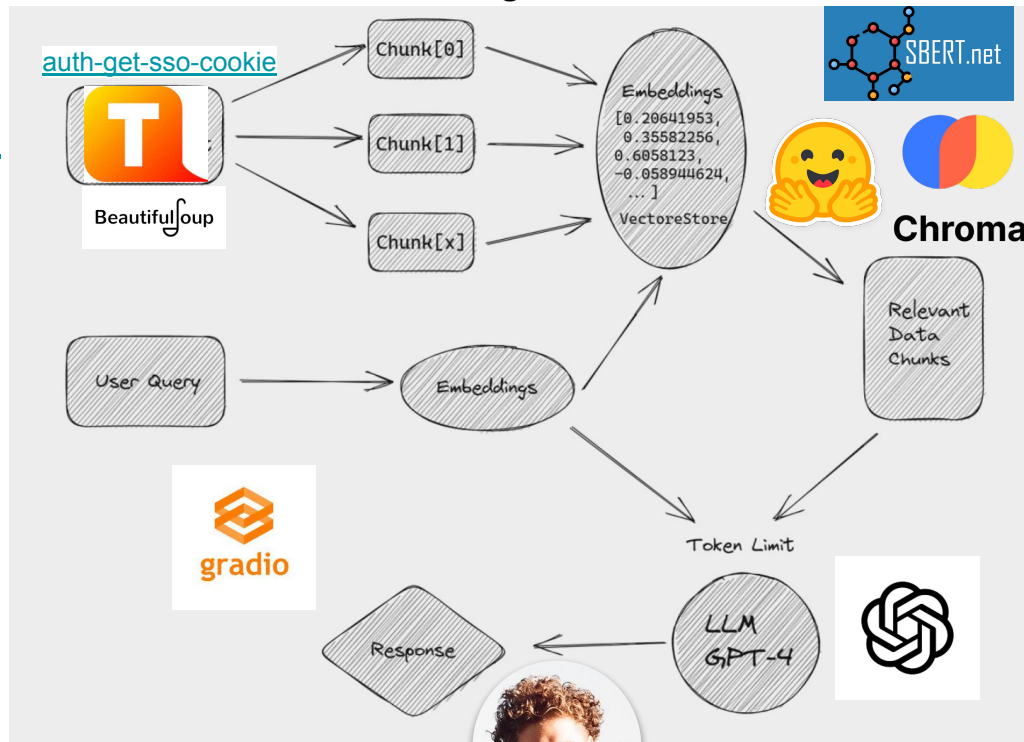


GF

chATLAS An AI Assistant for the ATLAS Collaboration



LangChain -



*Name credit: Sam Van Stroud
Thanks to M. Hart, N. Pond, J. Barr, A. Vaitkus for the gentle ribbing
IML Meeting, 9th April 2024

Simple to start



LangChain -

auth-get-ss0-cookie

Chunk[0]



How do I calculate the z_0 of a track with respect to a vertex?

You can calculate the z_0 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());
```

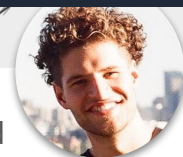
where $\text{trk.vz}()$ represents the point of reference for the z_0 calculation (in this case, the beamspot position along the z axis).

DM

chATLAS An AI Assistant for the ATLAS Collaboration

GF

*Name credit: Sam Van Stroud
Thanks to M. Hart, N. Pond, J. Barr, A. Vaitkus for the gentle ribbing
IML Meeting, 9th April 2024



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

Input query

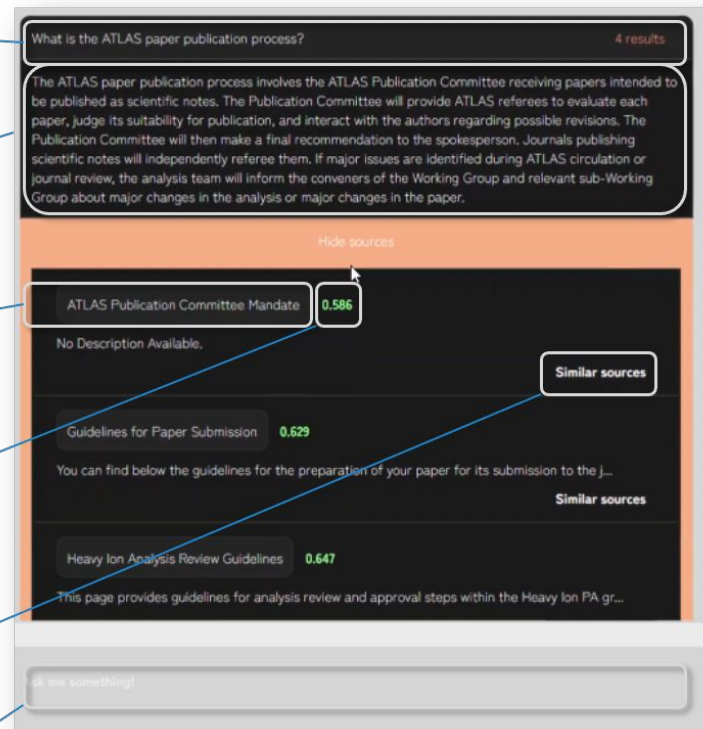
LLM response

Source title

Similarity score

Instant similarity search

Next query entry



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.

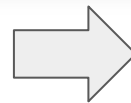
Hide sources

Source title	Similarity score
ATLAS Publication Committee Mandate	0.586
Guidelines for Paper Submission	0.629
Heavy Ion Analysis Review Guidelines	0.647

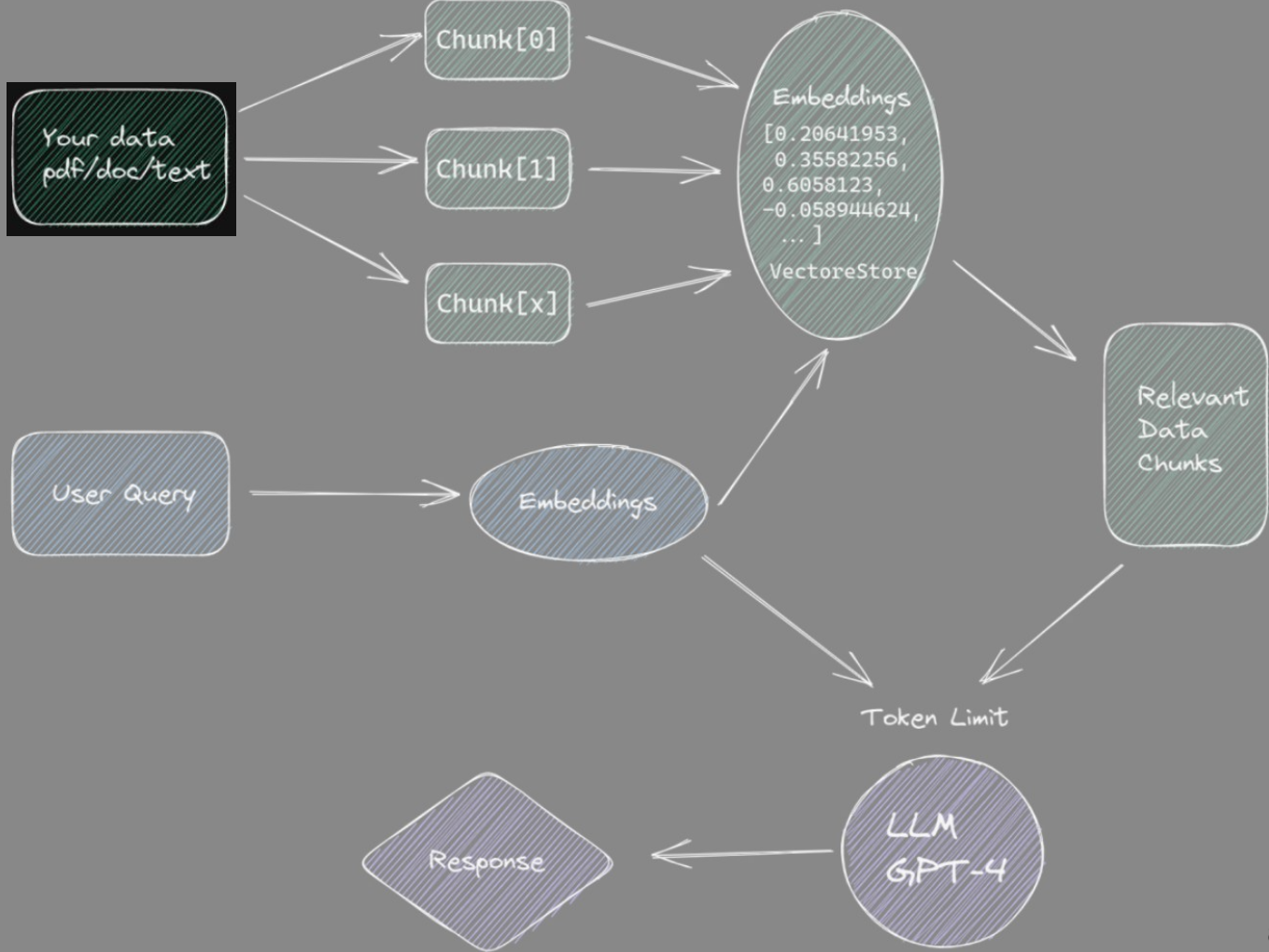
Similar sources

Similar sources

ask me something!



1. Scraping & Document Collection



Expanding footprint: + more TWikis +CDS



RL

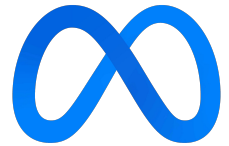
Significant expansion of collected dataset:

- 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



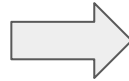
CR

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!

Atlas forums

<input type="checkbox"/> 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 other
Documentation and Communication	hn-atlas-docAndCommunication	This forum covers general announcements concerning Documentator
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 patch
Software Developers Announcements	hn-atlas-SWDevelopersAnnounce	Announcements, which are intended for developers of ATLAS offline
Category : Computing Offline Software (28)		
Architecture Team: Core Software Architecture and Design	hn-atlas-SWAArchitecture	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 I
Athena Event Display	hn-atlas-AthenaDisplay	Forum for help requests, suggestions, discussions and comments reg
Atlas Support	hn-atlas-atlas-support	Support for Atlas fast simulation users
Atlas Working Group	hn-atlas-atlas2-newdev-val	This forum is aimed for the development, validation and testing of the
Bugs	hn-atlas-Preimbugs	Atlas Preliminary/unconfirmed Bugs, Problems, Frustrations, Fixes
Bytestream Initiative	hn-atlas-bytestream-initiative	A finite-lifetime forum to coordinate bytestream infrastructure develop
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 req
Generator Validation	hn-atlas-generator-validation	Please report release specific errors and problems to the Forum "Rele
Forum for VPI support	hn-atlas-vpi-help	Common forum for MC generator responsables in ATLAS
New Job Configuration	hn-atlas-NewJobConfiguration	Forum for VPI (Virtual Point 1) support
Offline Commissioning	hn-atlas-offline-commissioning	In this forum we will discuss the migration to Configurables, JobPrope
		Offline Commissioning; For discussion about the offline software use

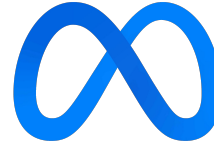


Code
(in alpha)

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



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

HEP C++ course

Based on the work of Sébastien Ponce

sebastien. Ponce

Foreword

What this course is not

- It is not for absolute beginners
- It is not for experts
- It is not complete at all (would need 3 weeks...)
 - although is it already too long for the time we have
 - 234 slides, 330 pages, 10s of exercises...

How I see it

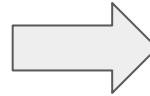
Adaptive: pick what you want

Interactive: tell me what to skip/insist on

Practical: let's spend time on real code

Where to find latest version ?

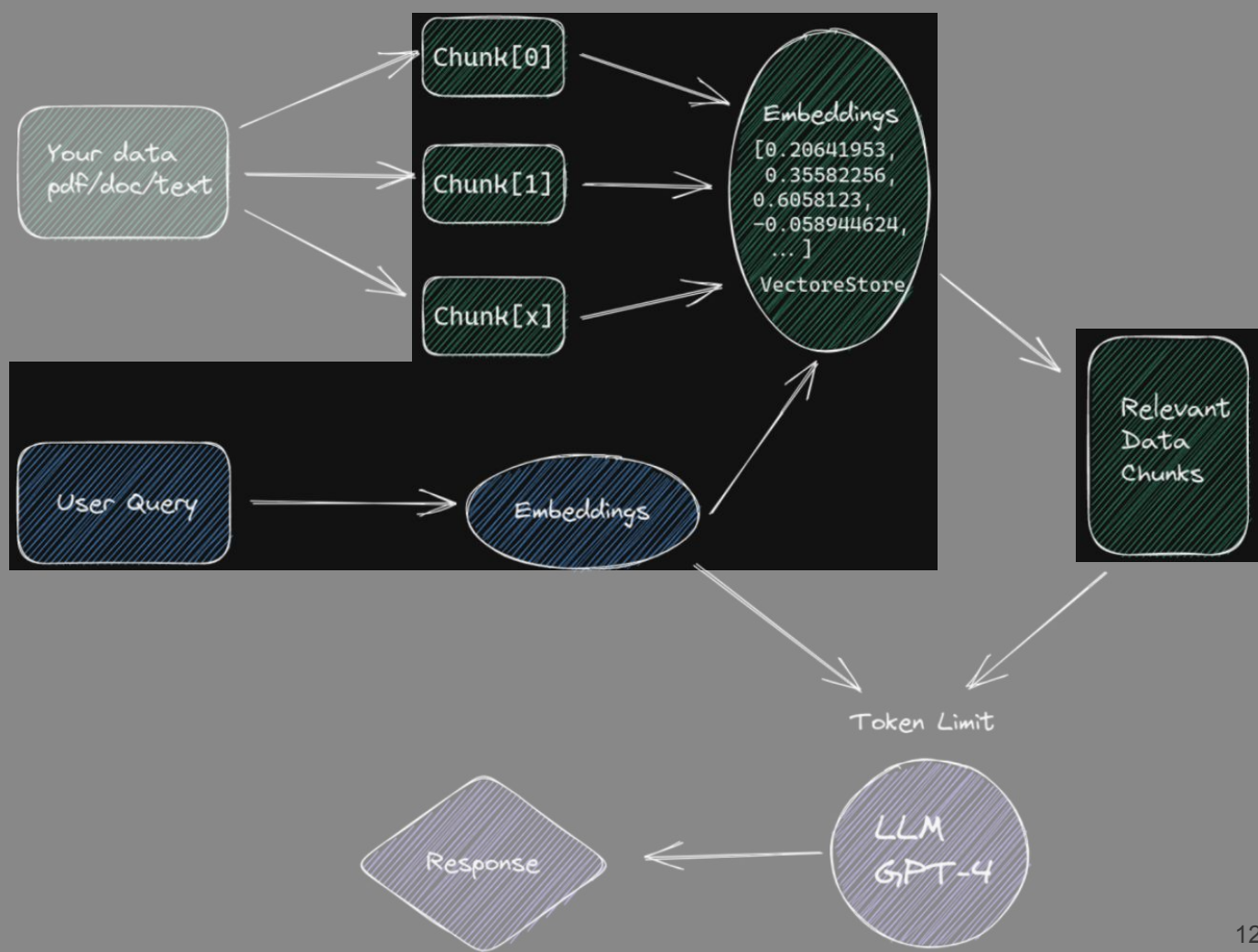
- pdf format at <http://cern.ch/sponce/C++Course>
- full sources at <https://github.com/hsf-training/cpluspluscourse>



Code
(In alpha)

```
Open file: C++Course.md
Based on the work of Sébastien Ponce sebastien.ponce@cern.ch
# More Foreword What This Course Is Not
It is not for absolute beginners It is not for experts It is not complete at all (would need 3 weeks...)
although is it already too long for the time we have 234 slides, 330 pages, 10s of exercises...
# How I See It Adaptive Pick What You Want
Interactive tell me what to skip/insist on
Practical let's spend time on real code
# Where To Find Latest Version ?
pdf format at http://cern.ch/sponce/C++Course full sources at https://github.com/hsf-training/cpluspluscourse
# Base Tool Do More More Courses
The HSF Software Training Center A set of course modules on more software engineering aspects prepared from within the HEP community https://hepsoftwarefoundation.org/training/curriculum.html
# Base Tool Do More Outline
```

2. Embedding & Vector Store



Embeddings



GF

Model dependent. Using: [sentence-transformers/all-MiniLM-L6-v2](https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2)

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

Search

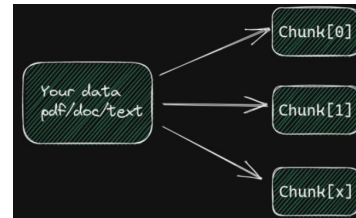


Chroma

RL introduced a child-parent structure:

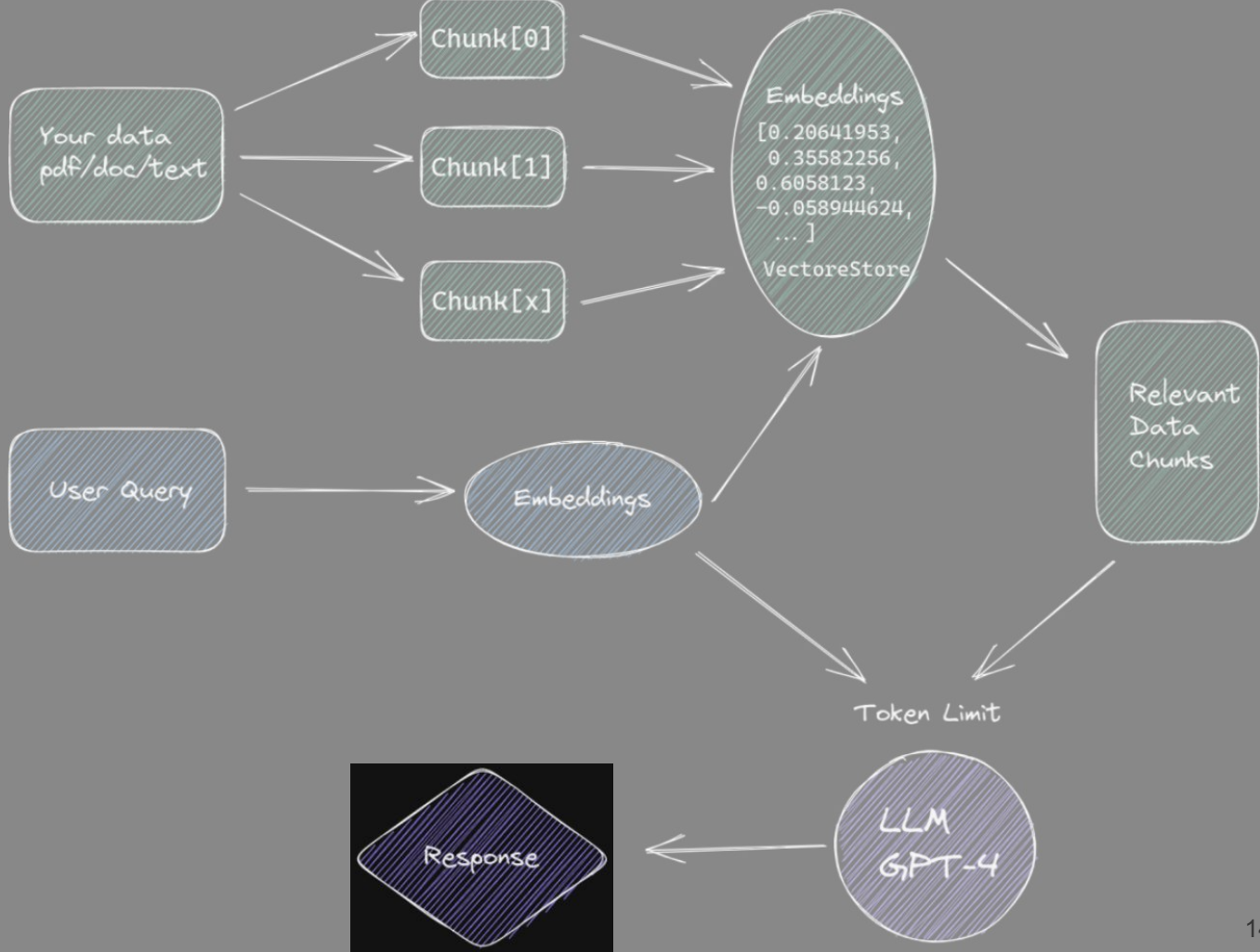
- **Parent** is full doc. **Children:** Chunks of ~200 tokens
- Search children, parent to LLM

Improved results on basic tests



RL

3. Evaluation



How good is it?

Quality control, test performance, test evolutions

JC got the work started with generating test data:

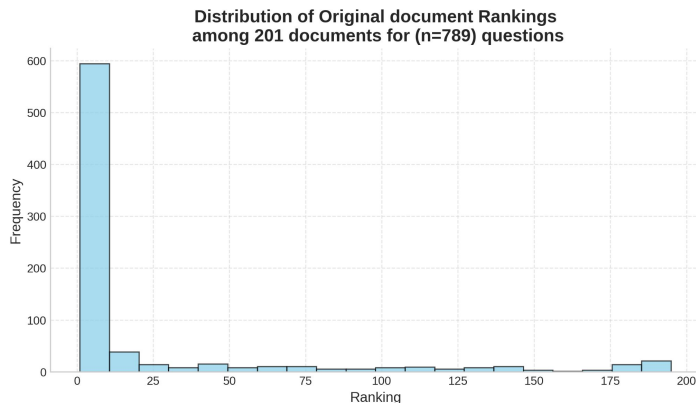
- For each TWiki, ask an LLM to list X questions that can be answered by the TWiki & response

Invert for testing:

- **Search:** Ask question and get the rank of the TWiki among the documents (see plot)
 - **With current setup, correct doc returned $\frac{2}{3}$ of time**
- **Assistant:** Ask question and see if answer matches (cosine sim) the “known” answer

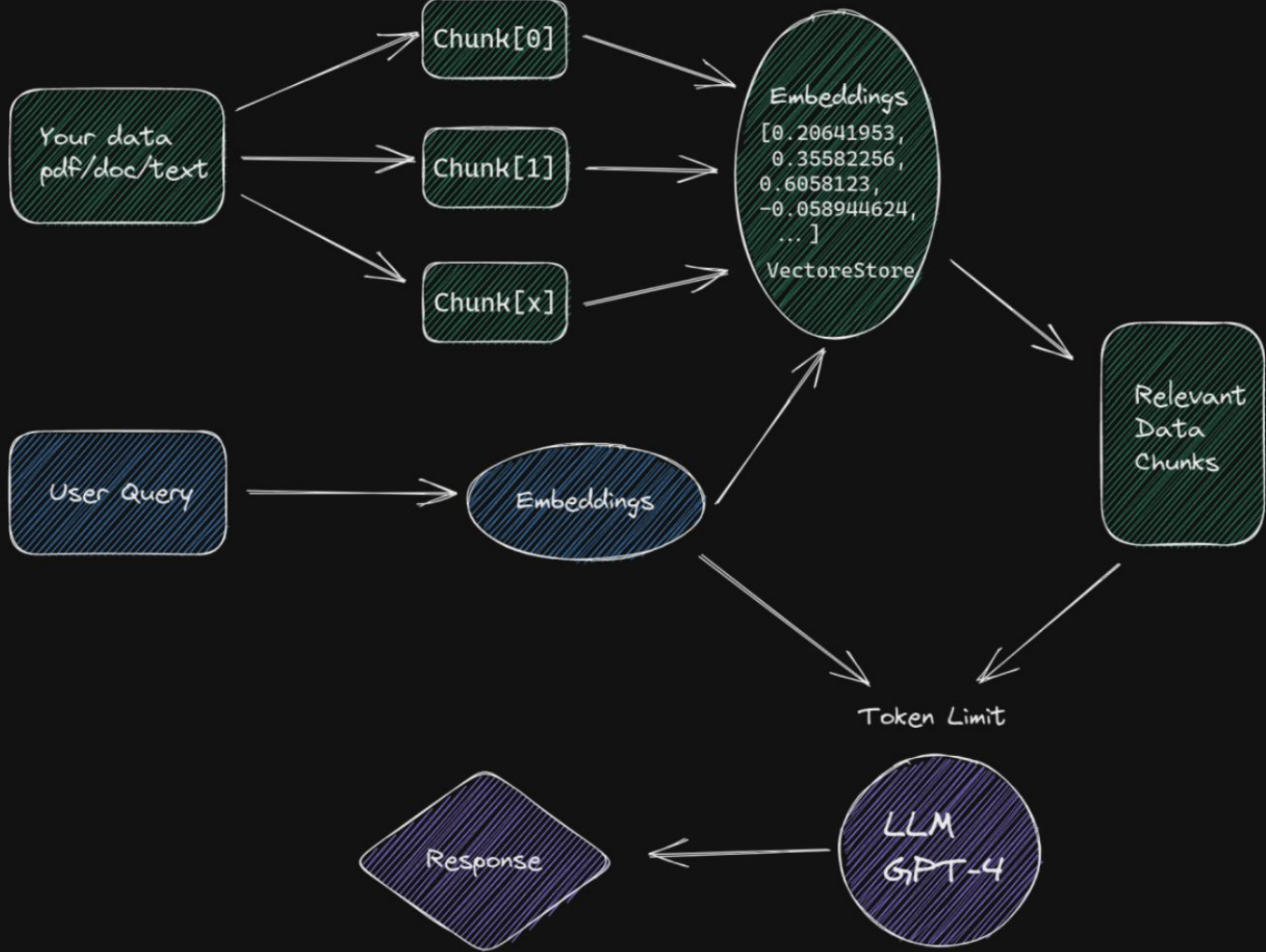


JC

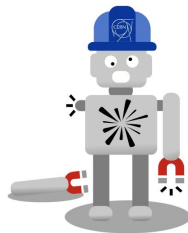


Small test: Ability to find correct TWiki using 789 questions from 201 TWikis

4. UI & Serving in Production

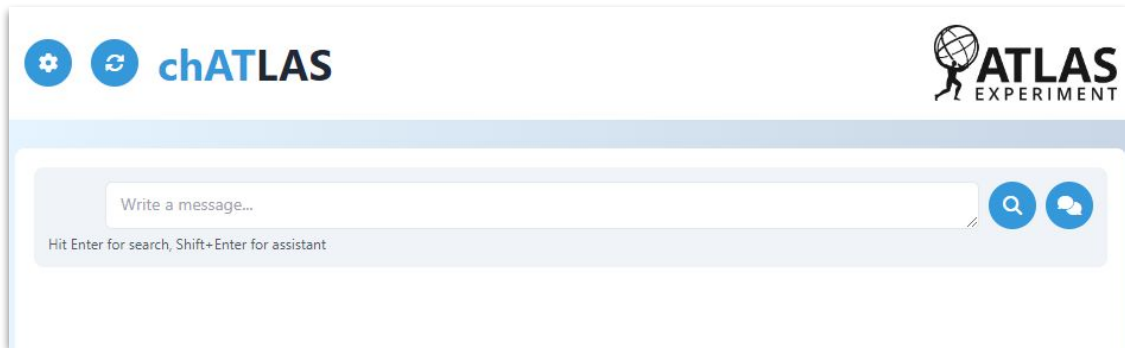


chATLAS in Production



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)



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 p...
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 hypotesi...
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 intr...
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

1. Incorporate more sources:

- CDS papers (in alpha), confs, pubs, INTs
- e-groups (scraped, to be embedded)
- 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

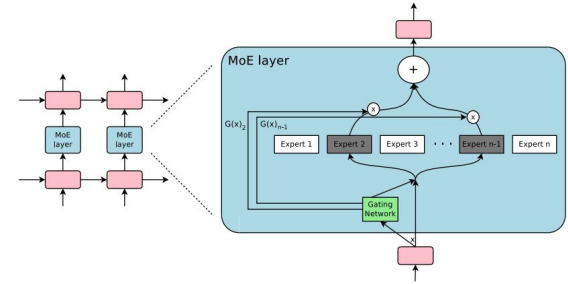
4. Fine tune of open source LLM a la [AstroLLaMA](#) - low stats show promise

5. Serving open source LLM via small API running at university

6. Improve quality through advanced RAG techniques

- 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 Ixplus if PaaS is unresponsive*
- [dataset locations](#)

Conclusions

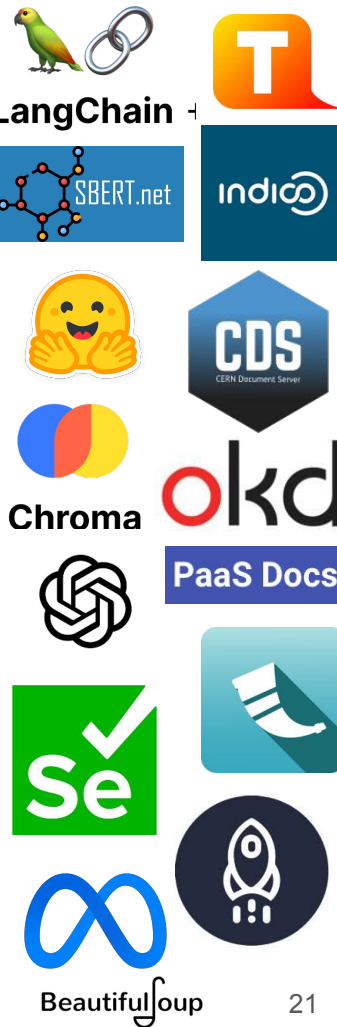
Finding information inside ATLAS is hard. Made efficient, collectively we can achieve more and potentially discover new physics sooner.

- 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

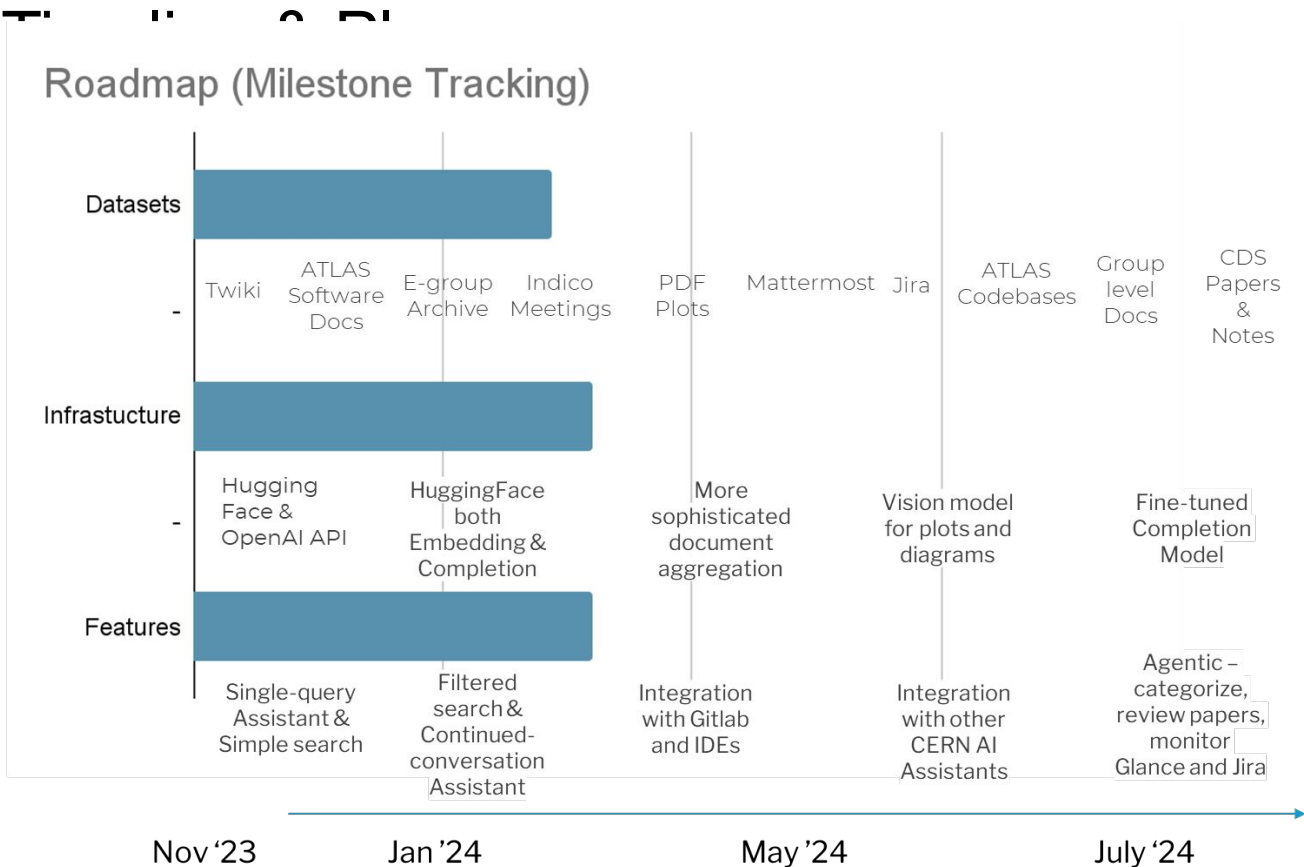
Task	DB	Twiki	ATLAS Software Docs	E-group Archive	Indico Meetings	PDF Plots	Mattermost	Jira	ATLAS Codebases	Group level Docs	CDS Papers & Notes
Scrape		2k+	500+	10k+	~1k+	~1k+					~5k+
Convert											
Chunk & Embed											

Not yet started

In Progress

Complete

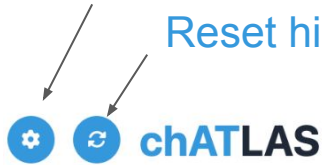
Specific



Overview of App

Configure (see next slide for options)

Reset history

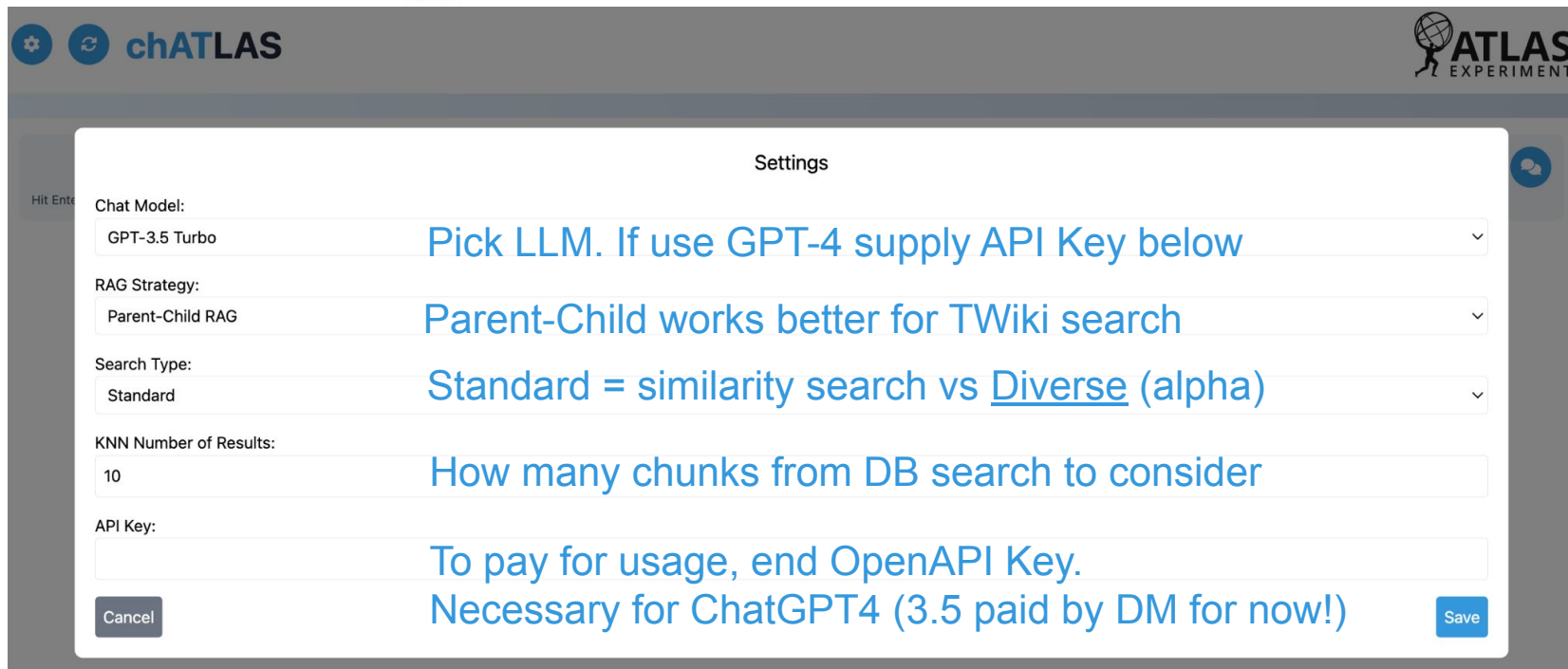


Enter = search mode. Lists Twikis (with links) in order of search strategy (see configurations)

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 



Settings

Hit Enter

Chat Model: GPT-3.5 Turbo **Pick LLM. If use GPT-4 supply API Key below**

RAG Strategy: Parent-Child RAG **Parent-Child works better for TWiki search**

Search Type: Standard **Standard = similarity search vs Diverse (alpha)**

KNN Number of Results: 10 **How many chunks from DB search to consider**

API Key: **To pay for usage, end OpenAPI Key. Necessary for ChatGPT4 (3.5 paid by DM for now!)**

Cancel Save