

# A2RCHI

## Building Intelligent Retrieval Systems for SUBMIT

**Mariarosaria D'Alfonso**, Pietro Lugato, Luca Lavezzo, Dmytro Kovalskyi, Zhangqier Wang, Christoph Paus,  
Julius Heitkoetter\*, Ivan Paus\*, Ludovico Mori \* (CMS/Physics@MIT)  
Tim Kraska, Matthew Russo, Jason Mohoney (CSAIL@MIT)

Submit workshop – 23 January 2026

(\*) undergrads that moved to other projects/fields

# A2RCHI: AI-Augmented Research Chat Intelligence



A2RCHI is an open source, end-to-end RAG framework

- Provides large language model (LLM) based services, with access to specific knowledge relevant to whatever use case.
- Augments knowledge via use of Retrieval Augmented Generation (RAG).
- Customizable AI solution for Education, Operation & Research.

# Retrieval-Augmented Generation (RAG)



## Why RAG?

- Base LLMs are trained on broad datasets, but may not contain your private or domain-specific data.
- **RAG** enhances LLMs by retrieving relevant documents and injecting them into the model's context before answering.

## In A2RCHI:

- RAG is used to combine LLM power with domain-specific documents, code snippets, and configuration guides

→ *This ensures responses are grounded in submit-specific knowledge.*

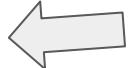
1. User Question  
↓
2. Search Vector DB (Top-K matches)  
↓
3. Add Context to Prompt, use chat history  
↓
4. Send to LLM  
↓
5. Generate Answer  
↓
6. Return to User

# Compatibility with different LLMs



*ready to use, can outperform in some area*

OpenAI GPT



Ollama



Lang Chain

*flexibility, cost savings, and customization,  
enhanced security/control*



LLama



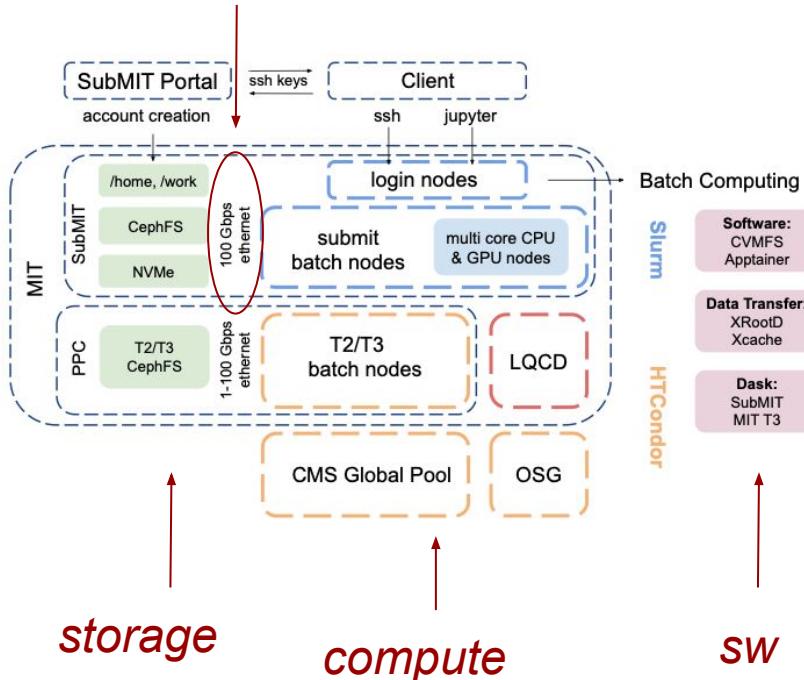
**Hugging Face**



Mistral

# Pilot projects

*fast network*



**subMIT** very flexible system,  
users utilize the sw/hw according to their requirements

## LLM project

→ help experts to support ticket handling  
→ chat available for users to ask questions



# Pilot projects

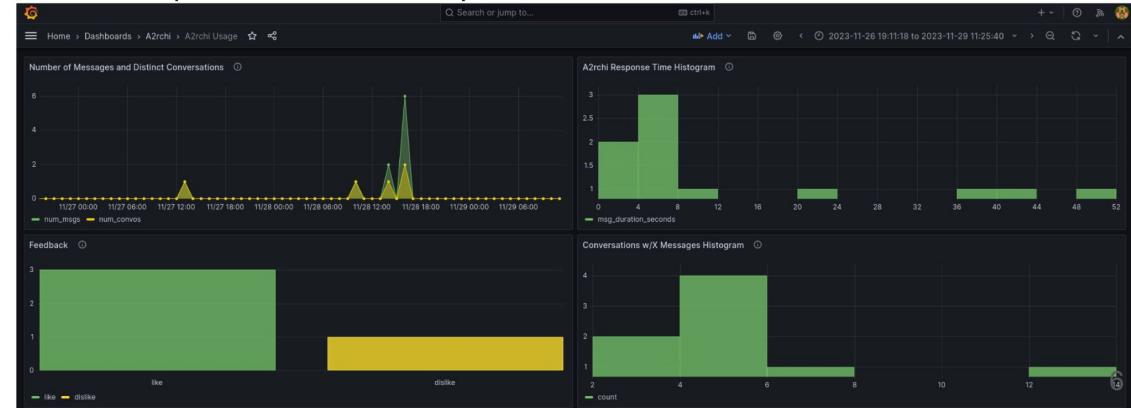


Running inference on a **dedicated HW**: i.e. 2 x Nvidia Tesla v100s

<http://submit75.mit.edu:7861/>

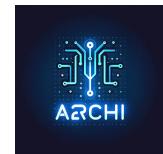
**Running local Ollama server**, an open-source platform focused on deploying and managing LLMs locally on organizational hardware. This approach enhances data privacy, cost efficiency, and performance compared to cloud-based services.

Access to **plot the various metrics** through Grafana





# Chat Service

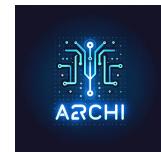


## Access

- Provide immediate 24/7 and accurate answers w/o the wait times with human agents
- Be direct, ask for sources, your history is preserved
- Easy escalation to human to follow up



# Where it get informations



Trained on Users Guide GitHub Examples

Anonymized Past Ticket to [submit-help@mit.edu](mailto:submit-help@mit.edu)

outdated material removed, i.e. sources are constantly scraped

specific to the project

past tickets, Q&A,  
users guides ...



Jira

Authentication  
SSO



webcrawler,  
git scraping



Files type





# Example #1



Can you list the options available on subMIT to install my collaboration-wide software and custom made libraries and packages ?



On subMIT, you have several options available for installing your collaboration-wide software and custom made libraries or packages:



1. Native system: The subMIT machines come with several pre-installed software such as python, C++, Java, MATLAB etc. that you can use.
2. Package and Environment Managers: You can use Conda, a package and environment manager, to install software. It could be an ideal choice for installing your custom made libraries or packages. Another package manager, spack, is also available.
3. Containers: You can use singularity or podman to create containers, which can be used to set up more complicated environments. This could be helpful for installing your collaboration-wide software.
4. CVMFS: Provided by CERN, this service can be used to set up many environments, which could be helpful in dealing with your custom libraries and collaboration software.

For setting up your software or environments, you can choose the most suitable one according to your needs. Please note that it's always preferred to use Conda if possible.

Enter your question for A2rchi





# Example #2



I'm external to MIT, but I collaborate with MIT researcher. How do I get an account on submit ? Please outline all the steps



To obtain an account on MIT's "submit" system as an external collaborator, follow these steps:



#### 1. Request via MIT Collaborator

- Your MIT researcher must contact the [MIT IS&T Research Computing team](#) or the relevant department administrator to initiate the request.
- Provide your full name, institutional affiliation, and purpose for access.

#### 2. Formal Invitation

- The MIT sponsor will receive a temporary invitation link to share with you (valid for 7 days).
- Click the link to create an MIT "Temporary" account ([Temporary Accounts - IS&T](#)).

#### 3. Identity Verification

- Complete identity verification using your external institution's credentials (e.g., university or employer email/SMS).

#### 4. Access Approval

- After account creation, your MIT collaborator must manually grant "submit" access through the [MIT Research Computing portal](#) (login required).

#### 5. Login to Submit

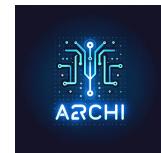
- Use your temporary MIT account credentials at [submit.mit.edu](http://submit.mit.edu) (if this is the target system).

For urgent access, your MIT collaborator should email [help@mit.edu](mailto:help@mit.edu) with subject "External Collaborator Access Request" and include your details and justification.

Note: Policies may vary by lab/dept—confirm with your MIT contact for lab-specific procedures.



# Rate your answers



# Pilot projects – Classrooms



**Education 2023-24:** 8.01, 8.511

- given relatively small documents or multiple full-size textbooks
- Supports Piazza Q&A and tutoring — reduces TA load while maintaining academic integrity

⇒ not only FAQs but more help with PSET and learning exercises



## Challenges being addressed

### content evolution

past knowledge still relevant, but outdated can lead to misleading answers

### weighted response

⇒ direct experts answers can be more useful than general material  
⇒ new information added this year should be relevant than the previous years

### misuse

avoid the students get the full solution

# A2rchi beyond submit



A2rchi is going to help also CMS experiment in some operational aspect:

where

**Tier0**

**Data Management ,  
Production and  
Reprocessing**

**DataQualityMonitoring  
shift**

how

**Chatbot** to facilitate onboarding of new team members,  
to help CMS shifter during operations

**Assistant** for meetings, JIRA tickets, and Mattermost,  
when giving tutorial

**Analytics** to spot area where we need to improve

**Operator** as agent to replace humans

**Alert system** to spot anomalies in monitoring system



# If you want to join the LLM effort !



<https://mit-submit.github.io/A2rchi/>

developers team:

a2rchi@mit.edu

support team:

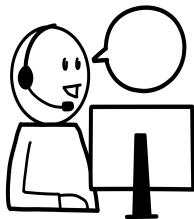
a2rchi-help@mit.edu

# BACKUP

# Interfaces and API/integration



Helpdesk



API calls in place to messaging services



Integration with CERN Ecosystem, driven by the ongoing projects

# Data sources



Relevant information to feed into each query

specific to the project  
past tickets, Q&A,  
users guides ...



*Jira*

Authentication  
SSO



webcrawler,  
git scraping



Files type

