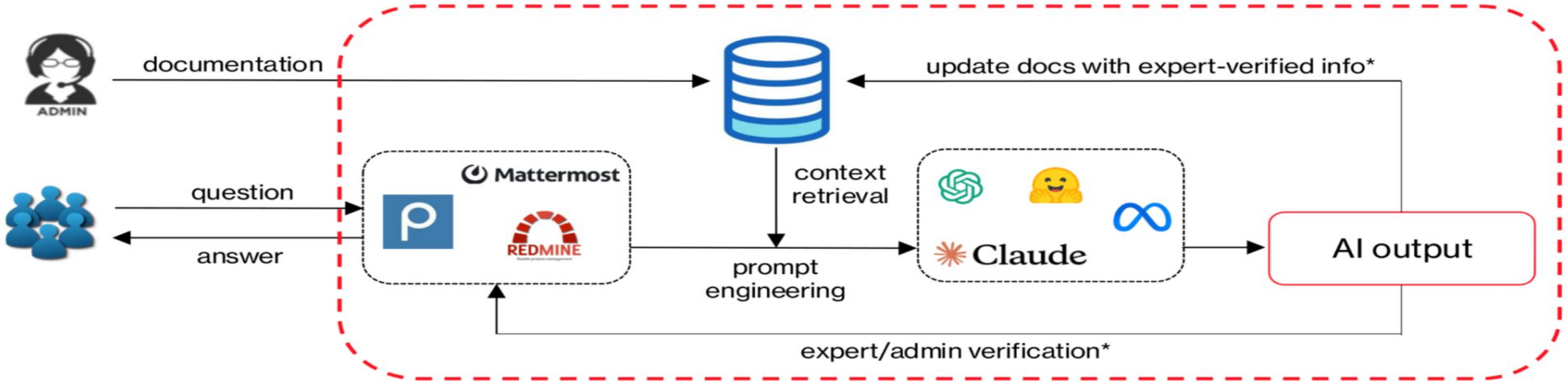
A2rchi On SubMIT

A quick look at how it works, what its being used for at submit and some ongoing work/plans

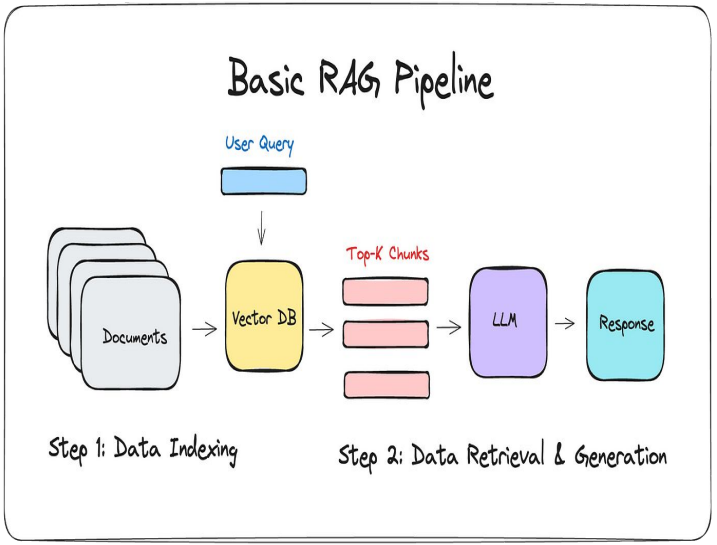
What is A2rchi?

- A2rchi is a RAG framework that supports multiple different services and pipelines for organizing an AI workflow
- Using your own inputted documents or webpages, A2rchi can compile relevant information into a database which it can then use to improve its output for question and Answer routines



Explanation of RAG

- RAG, or retrieval augmented generation is a technique for Artificial intelligence which involves adding more context to a given prompt for more accurate answers and better performance
- Generally, standard LLMs predict the next most common word in an answer to a users provided question based on their model weights
- By retrieving more specific context and adding it to a users query, the LLM can produce better and more helpful answers to a users question

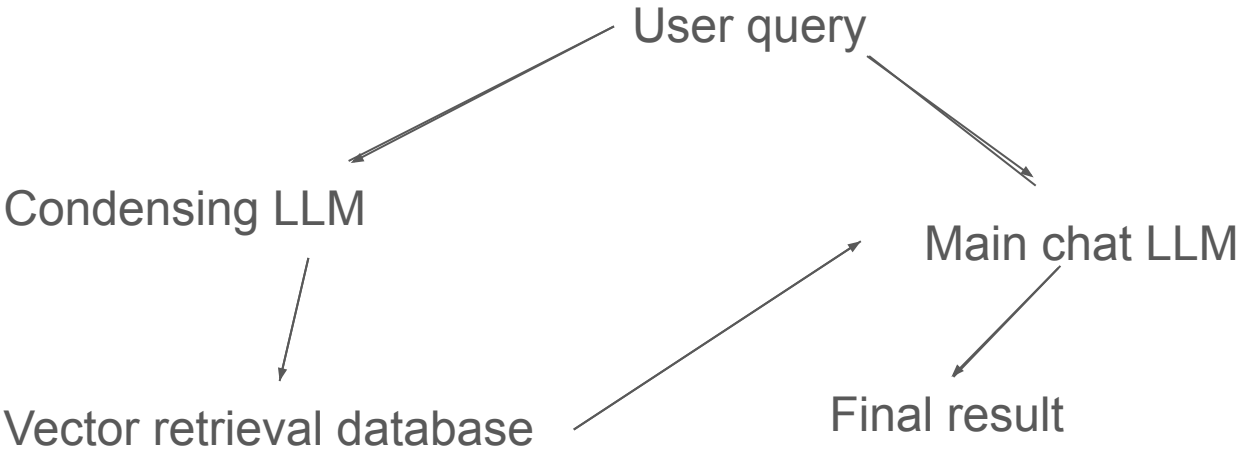


How does A2rchi specifically do this?

- Currently, A2rchi supports two main modes of information retrieval
  - Semantic similarity through a vector database
  - Lexical similarity through a text index
- Upon passing in documents, the textual information is processed and split into ‘chunks’
- Information in these chunks is then embedded and inserted into a vector database as well as a lexical search index
- Later the users question is translated to form a query to these two indices in order to find relevant information inside of them and use it to generate better context for the prompt passed to the LLM agent

Features of A2rchi as a framework

- Using the A2rchi framework, users are capable of defining their own pipelines for data flow in just a few lines of python code
- By then inputting a given dataset, it’s possible to customize A2rchi to fit one’s own needs
- For instance, our base pipeline looks a little something like this:



What is A2rchi used for on SubMIT

- Currently A2rchi has two instances used for submit:
  - one as a simple chatbot for submit users (trained on Submit information and past tickets)
  - One on redmine (trained on anonymized past tickets)
- For now Users can use it to easily find the answer to questions that a2rchi is trained on
- Feel free to access the chatbot version currently running at

  http://submit75.mit.edu:7861

Possible Future plans and applications for A2rchi

- Currently the A2rchi team is working with CMS to deploy it for use in scientific environments
- Possible applications include
  - A helper buddy for shifters monitoring the experiments
  - A retrieval agent for analysis information
  - A dedicated analysis code/framework generation tool

How to setup and use A2rchi

- Currently A2rchi is open source and can be found at the following github repo: https://github.com/mit-submit/A2rchi
- To use A2rchi the clone into the repository and pip install it into your python environment of choice
- Ensure that you then have either docker or podman installed on your machine
- Then simply run the following command in the command line:

A2rchi create –name <name for the a2rchi instance> –config <path to configuration file> –podman (if you have podman installed)

END OF PRESENTATION