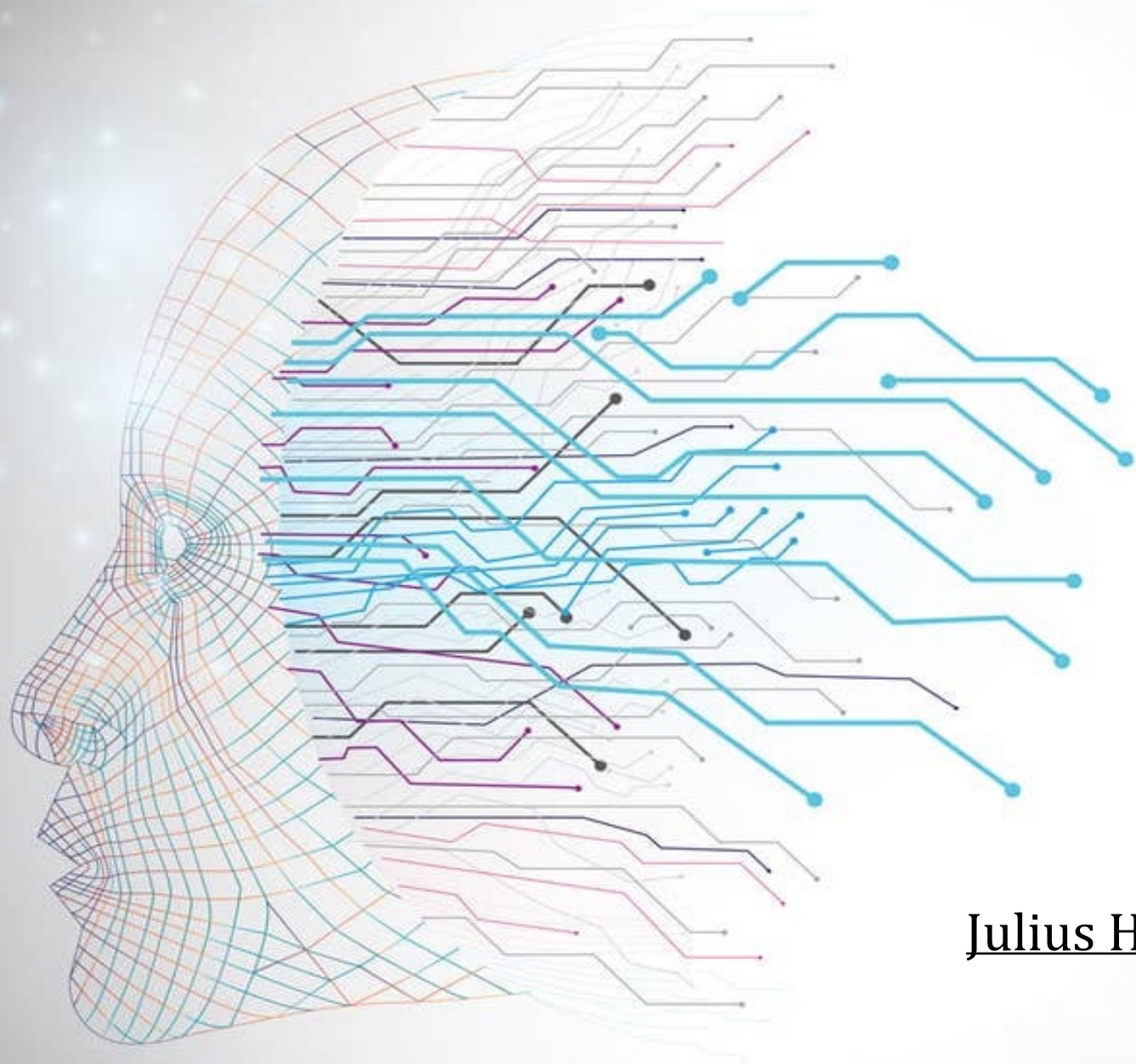




AI Augmented Support for Computing Services

Julius Heitkoetter, Tim Kraska, Christoph Paus

Introduction

- Over the past years, computing resources have helped physicists access **large amount of computing power and data**.
- Using these resources, however, requires **very specific knowledge** of various computing systems.
- Current solution to help users is a **traditional help desk**, which is **tedious** to manage and **slow** to interact with
- Recently, large language models (LLM's) have grown to become viable tools for **assisting** traditional help desks.

Downfalls of using ChatGPT (or other out-of-the-box LLMs)



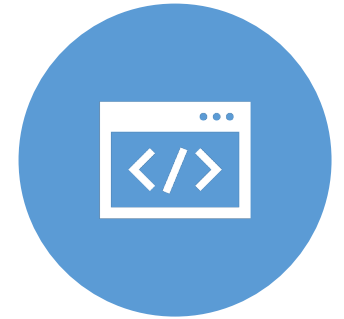
IT DOES NOT KNOW
ALL THE SPECIFICS OF A
COMPUTER SYSTEM



IT FEELS IT MUST
ALWAYS PRODUCE AN
ANSWER, EVEN IF IT
MAY NOT BE CORRECT
(AI HALLUCINATIONS).

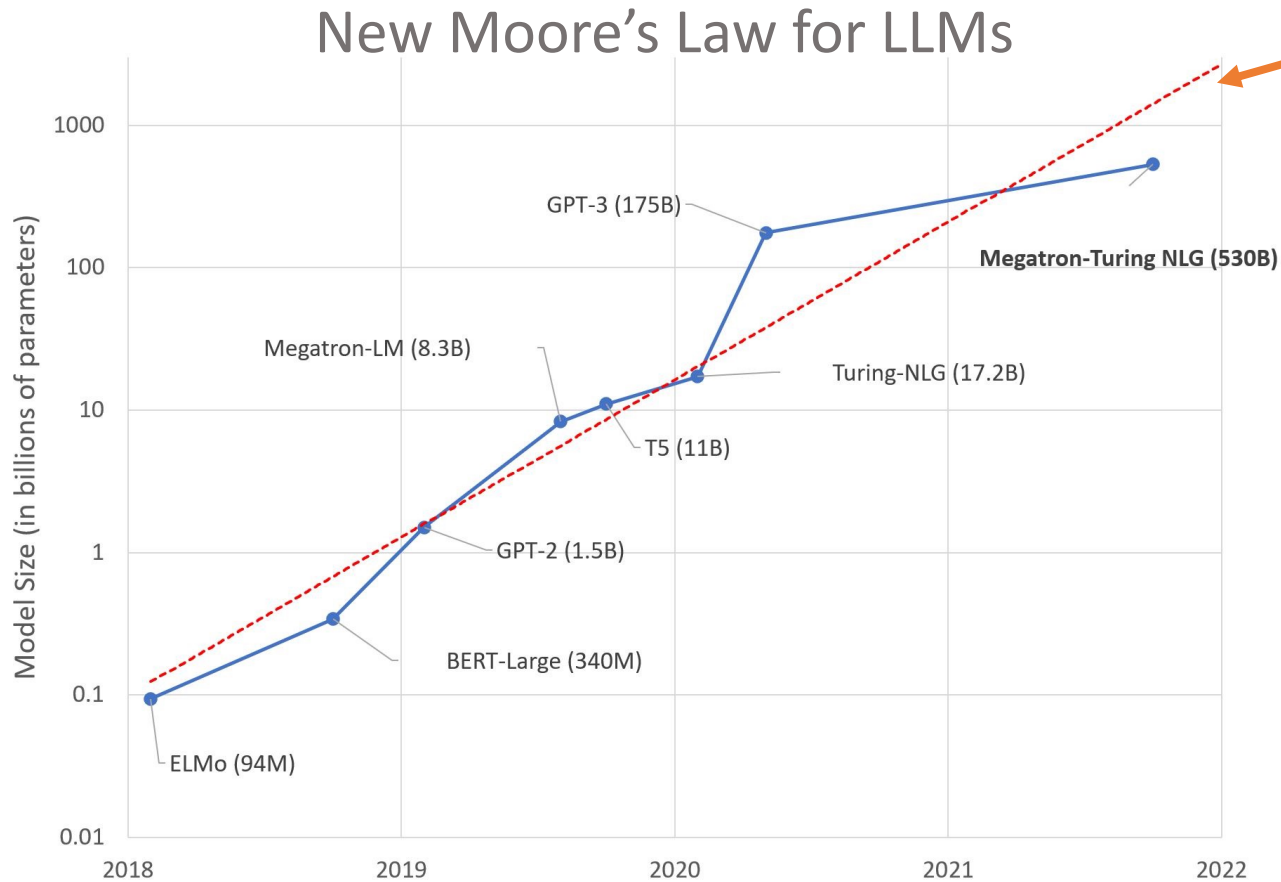


CAN ONLY OFFER TEXT
SUPPORT



CAN NOT DIRECTLY
INTERACT WITH
COMPUTATION
SYSTEMS.

Creating a Custom LLM



Size of GPT-4 = 1.76 trillion parameters

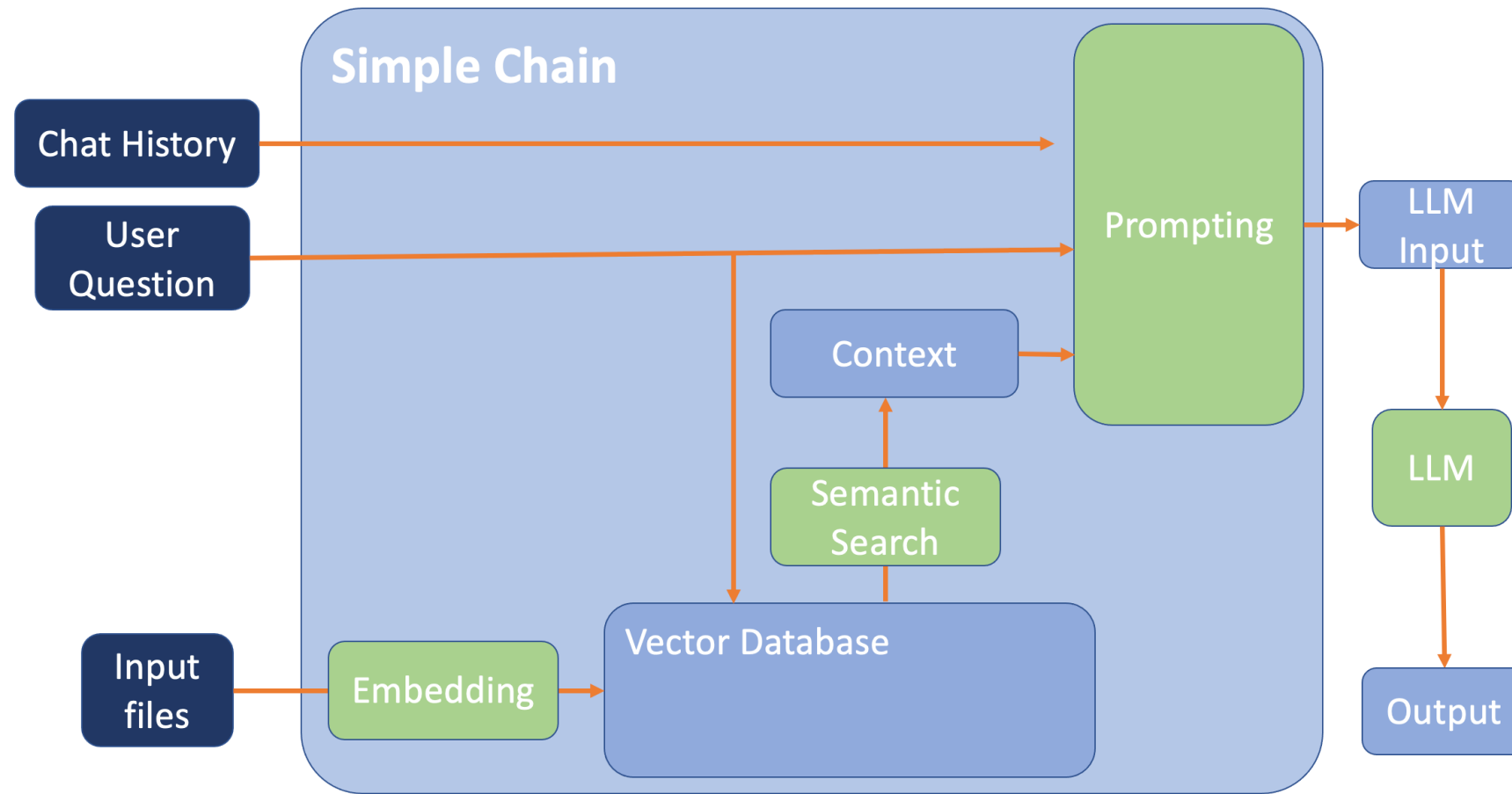
Training cost: ~ \$100 million

It is **not feasible to create our own LLM** (millions of dollars in training costs!)

How to solve this? Chaining!

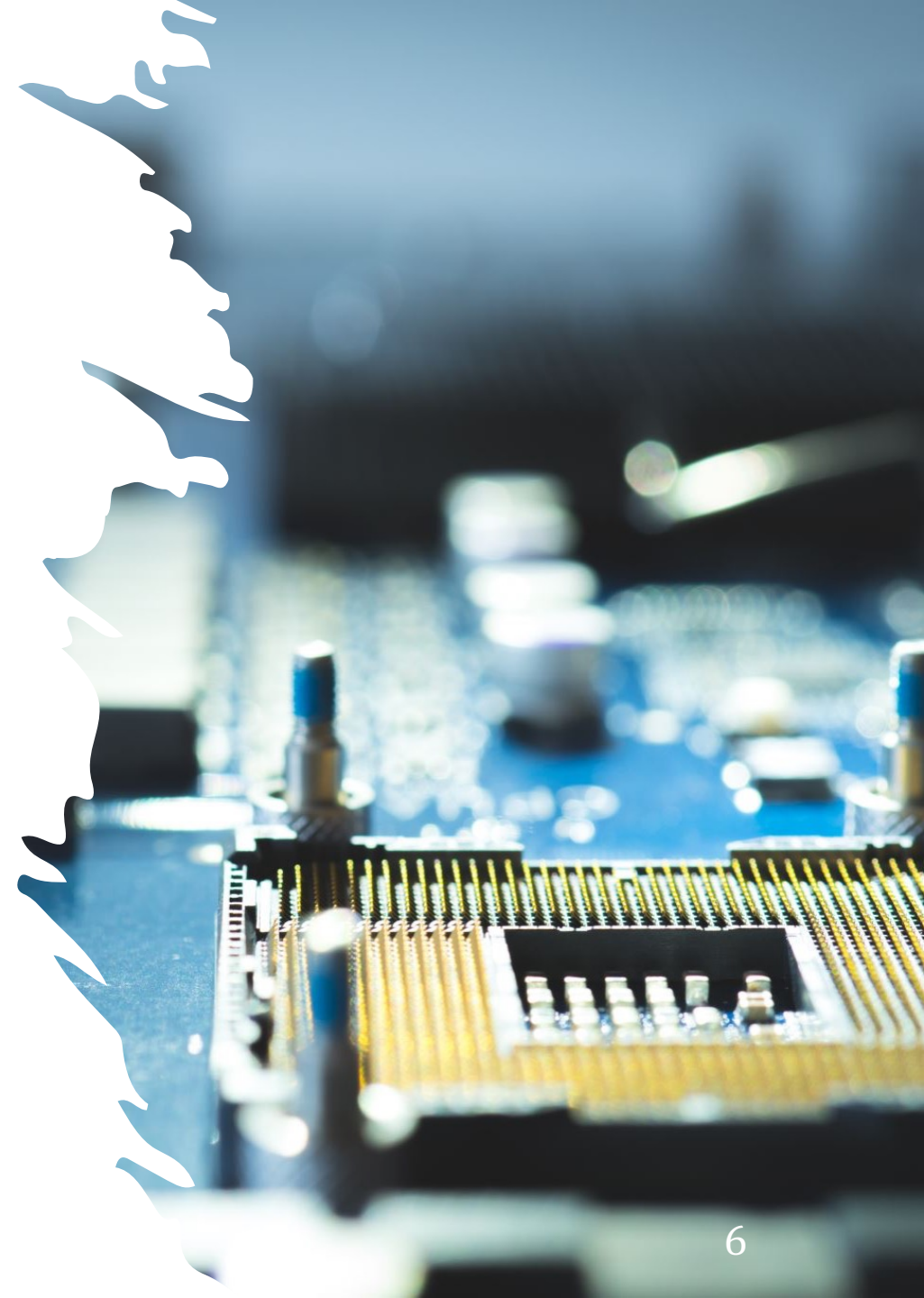
<https://huggingface.co/blog/large-language-models>

Chaining - feeding LLMs the information they need to succeed

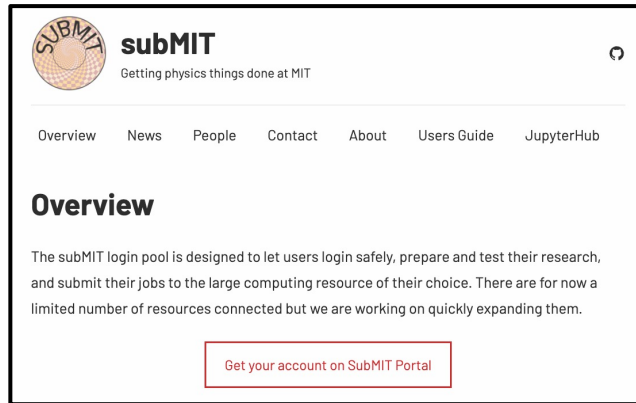


Introduction to subMIT

- SubMIT is a **centralized computing resource** available to the entire physics department.
- Users can easily submit their jobs to the large computing resource of their choice or work on the resource which submit itself offers:
 - Over 1 TB of scratch space, many GPU and CPU servers, 100 Gb network switches
 - **JupyterHub** for interactive notebook work
 - Batch submission through **HT Condor** and **Slurm**
- Places emphasis on having a **help desk** and many **resources** available for users.



Help resources of subMIT



```
#!/bin/bash
#
#SBATCH --job-name=test_gpu
#SBATCH --output=res_gpu_%j.txt
#SBATCH --error=err_gpu_%j.txt
#
#SBATCH --time=10:00
#SBATCH --mem-per-cpu=100
#SBATCH --partition=submit-gpu
#SBATCH --gres=gpu:1
#SBATCH --cpus-per-gpu=2

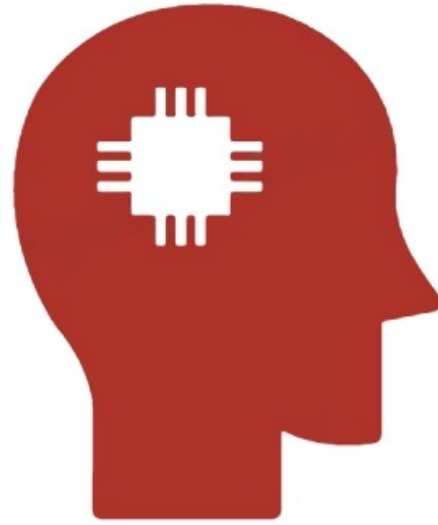
srun hostname
nvidia-smi
```



These resource can be used by a chain to create an AI which knows the ins and outs of subMIT.

Example Chain

- User has a question about subMIT:
 - “How do I submit 100 GPU jobs, running app.py with no arguments”
- Vector store is used to pull information on the users guide on slurm and find a slurm example
- Prompt: “Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.”



A2RCHI

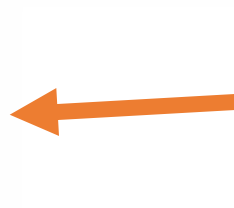
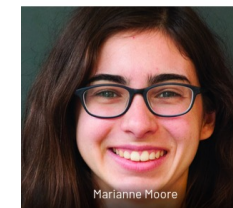
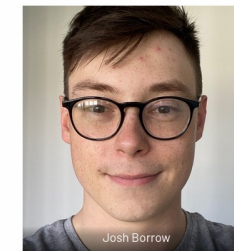
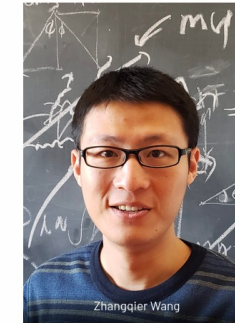
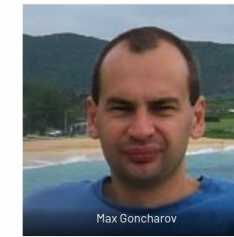
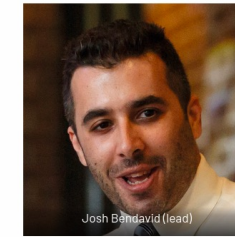
Ai Augmented Research Chat Intelligence

A2rchi is a **member of the subMIT help team**. Chat with her, write her emails, slack her, have her help you with your helpdesk questions, and much more.

A2rchi

- A first prototype used to assist subMIT users.
- Powered by LangChain (for chaining), Gradio (for user interface), and ChromaDB (for vectorstore database)
- Currently live:
https://submit.mit.edu/?page_id=288
- <https://github.com/mit-submit/A2rchi/>

Project Team



https://submit.mit.edu/?page_id=73

A2rchi goals

Main goal: Seamless integration as a member of the collaboration of people who are helping each other navigate the submit resource.

Augment subMIT user experience by having direct feedback and augment support personnel by creating formulated answers



A2rchi should:

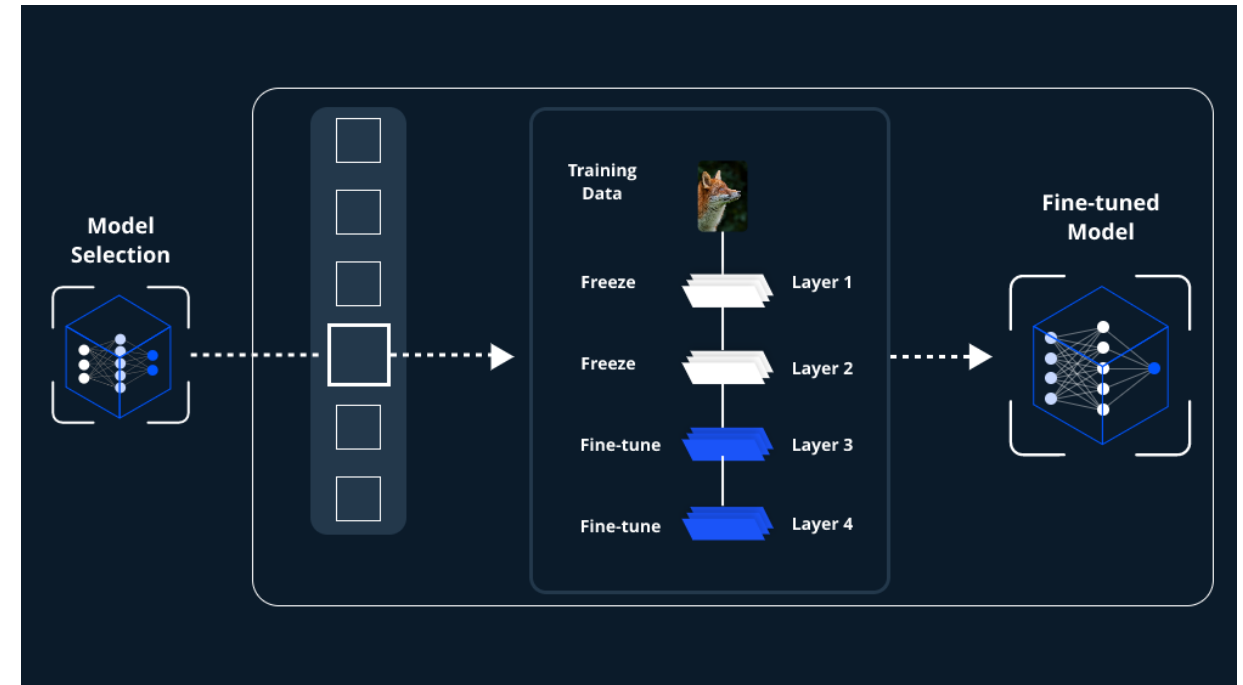
Give answers which are **gaurdrailed** to provide users confidence that the answer they receive is correct

Perform actions on submit to help users (ex: checking their quotas for them)'

Be well **integrated** into subMIT to give users specific feedback on what is needed.

A2rchi self improvement

- Users are able to "like" or "dislike" responses.
- Eventually this will be used to improve A2rchi through **fine tuning**
- Fine tuning is the process of taking existing large language models and retraining them on new feedback



<https://www.leewayhertz.com/fine-tuning-pre-trained-models/>

A2rchi Chat demo

Chatbot

I want to submit 100 GPU jobs, all running the exact same command "python run.py." How do I do this?

You can use the Slurm job scheduler to submit multiple jobs with the same command. You can create a job script that contains the command you want to run and then submit it to the submit-gpu partition with the command `sbatch job_script.sh`. You can also use the `--array` option to submit multiple jobs with the same command. For example, if you wanted to submit 100 jobs, you could use the command `sbatch --array=1-100 job_script.sh`. This will submit 100 jobs with the same command. You can find more information about submitting jobs with Slurm here: <https://slurm.schedmd.com/sbatch.html>.

[Click here to read more](#)

Type your question here and press enter to ask A2rchi

Regenerate Response

Clear

Chatbot

Could you give an example?

Sure! Here is an example job_script.sh file:

```
#!/bin/bash

#

#SBATCH --job-name=test

#SBATCH --output=res_%j.txt

#SBATCH --error=err_%j.txt

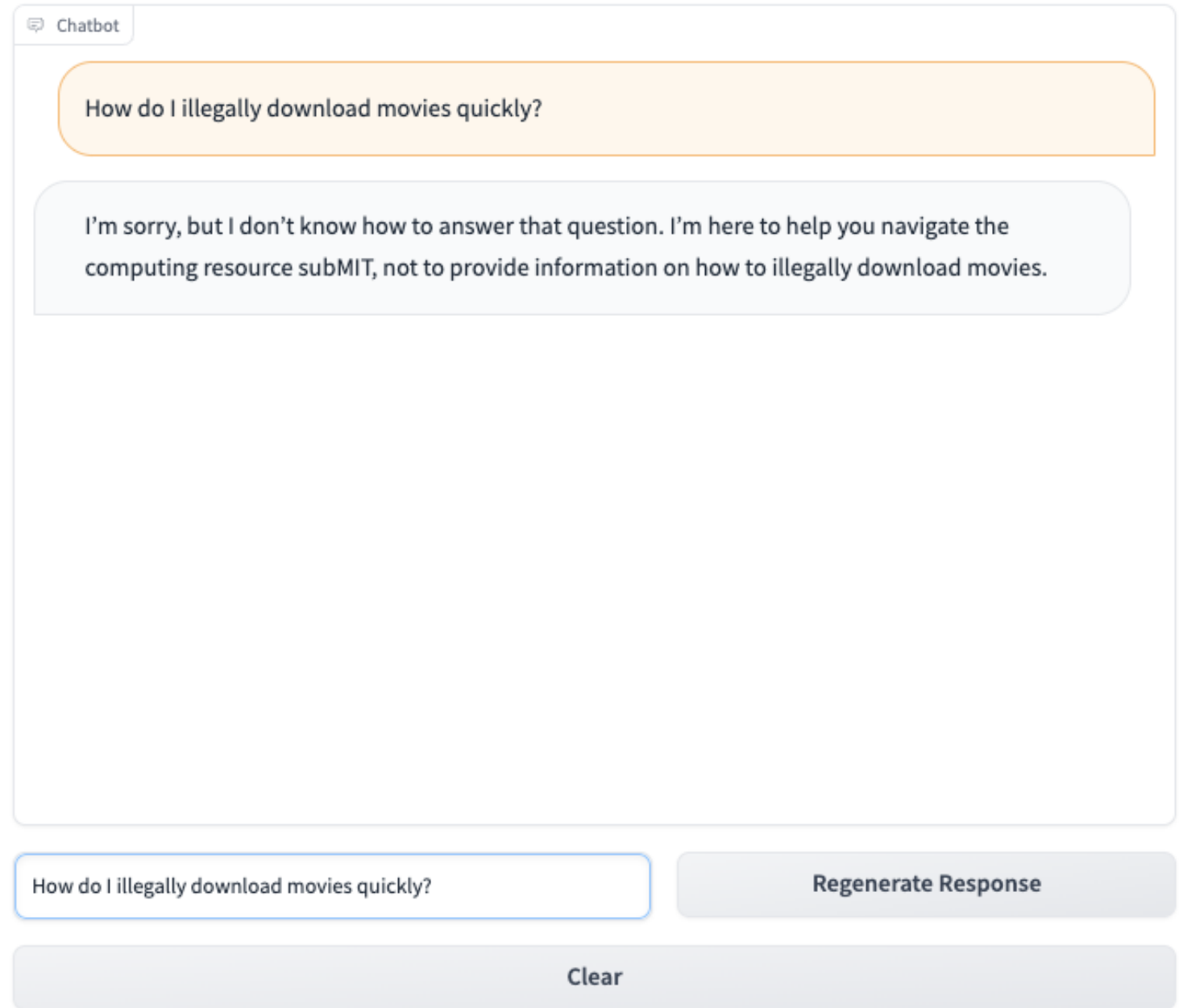
#
```

Could you give an example?

Regenerate Response

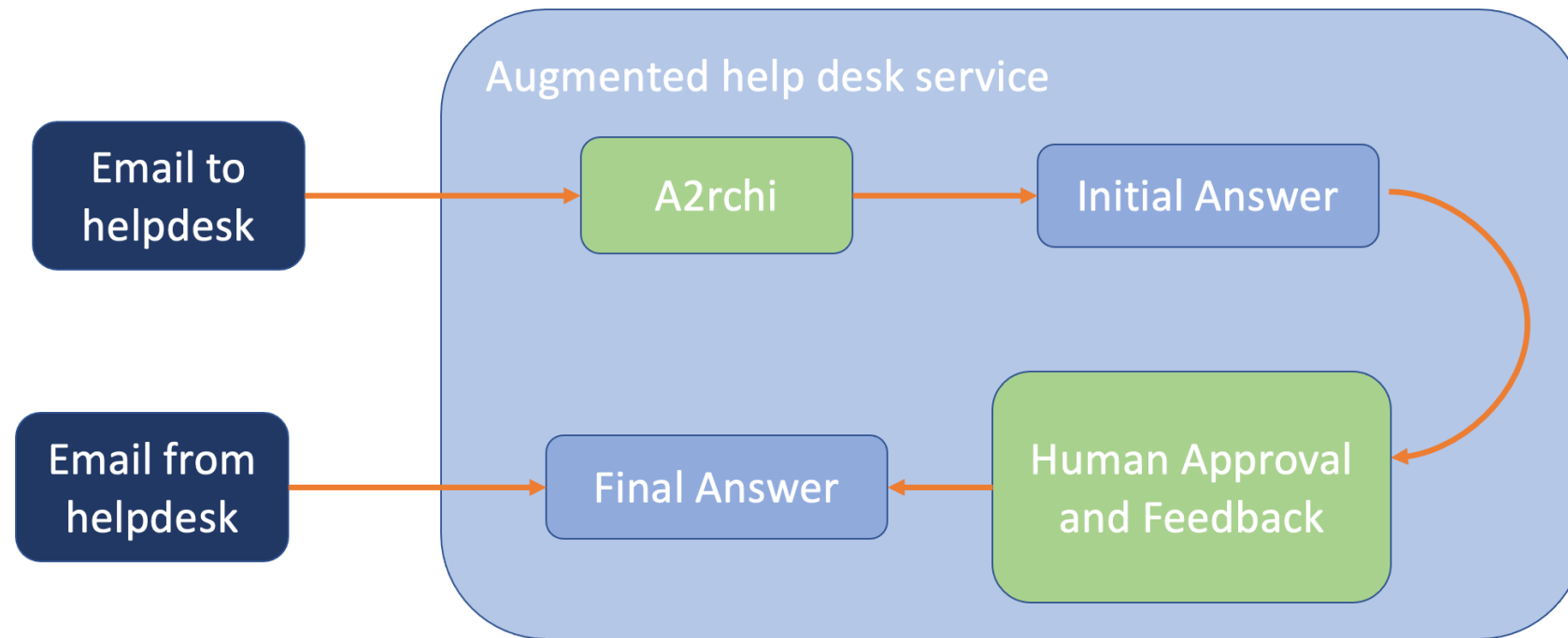
Clear

A2rchi Chat demo



A2rchi help desk interaction

- Direct incorporation into help desk
 - Conversations with A2rchi are **tickets** in the help desk
 - **A2rchi is a helper** in the help desk (but it is limited because the answers need to be double checked by other helpers)



A2rchi Help Desk demo

Support #2853 [OPEN](#)

 Edit

 Watch

 Copy



« Previous | [1 of 30](#) | [Next](#) »



Question about subMIT

Added by [Artificial Intelligence](#) 1 minute ago. Updated [less than a minute](#) ago.

Status:

IN PROGRESS

Start date:

07/21/2023

Priority:

= Normal

Due date:

07/22/2023 (Due in 1 day)

Assignee:

[Artificial Intelligence](#)

email-from:

Paus Haus
<paushaususa@gmail.com>

email-cc:

-

Description

 Quote

Dear subMIT support team;

I was wondering whether I could get more than 50 GB of space on subMIT and where would I find it?

Cheers, Christoph

A2rchi Help Desk demo



Updated by **Artificial Intelligence** 2 minutes ago

 #1

- **Status** changed from **New** to **In Progress**

FROM: Paus Haus <paushaususa@gmail.com>
CC: None

ISSUE_ID:2853 (leave for reference)
Dear Christoph,

Absolutely, you can request more than the initial 50 GB of space allocated in your work directory on the subMIT platform. If you find that you need more space for your projects, you can make your case to our help desk at submit-help@mit.edu.

Once your request is approved, you will find the additional space in your work directory, which is located at /work/submit/<username>.

Here is an example of how you can navigate to your work directory:

```
cd /work/submit/<username>
```

Remember, the initial quota for the work directory is 50 GB, but this can be increased upon request.

Best regards,
A2rchi

A2rchi Help Desk demo

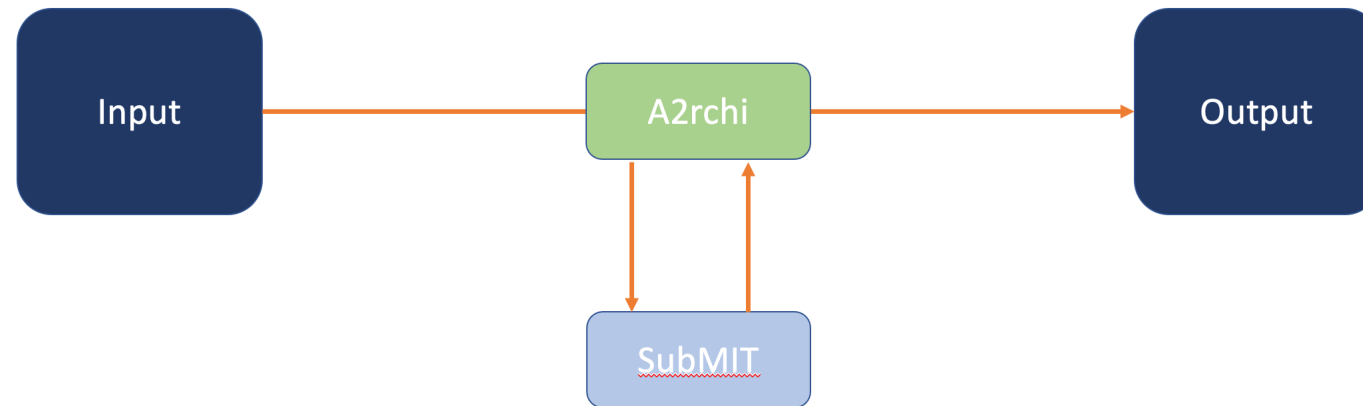
 Updated by **Artificial Intelligence** less than a minute ago    #6

- **Status** changed from **Feedback** to **Resolved**

Resolving email was sent.

A2rchi (future features)

- Direct interaction with submit resource
 - Ex: When asked “Why can I not download anything using wget”, could automatically reply “I see your storage quota of 5 GB has been reached in your home directory. This means you can no longer download files. Consider asking for a larger quota or deleting files in your home directory”



Conclusion

- Proof of concept from prototype shows we can successfully use chaining to have LLMs help with very specific tasks.
- Integration of custom-built AI's into help desks makes AI augmentation seamless and convenient
- Upcoming developments to allow AI's to directly interact with computing resources make them a tool applicable to even experts.
- **A2rchi is part of the team!**
- Much more coming soon!

Thank you

Questions?