INL/CON-24-77630-Revision-0

Exploring Large Language Models in the Cloud and On Premise

May 2024

Changing the World's Energy Future

Brandon S Biggs

Idaho National Laboratory

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/CON-24-77630-Revision-0

Exploring Large Language Models in the Cloud and On Premise

Brandon S Biggs

May 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

May 2024

Brandon Biggs

RMACC 2024

INL/CON-24-77630

Exploring Large Language Models in the Cloud and On Premise

Battelle Energy Alliance manages INL for the
U.S. Department of Energy's Office of Nuclear Energy

Idaho National Laboratory

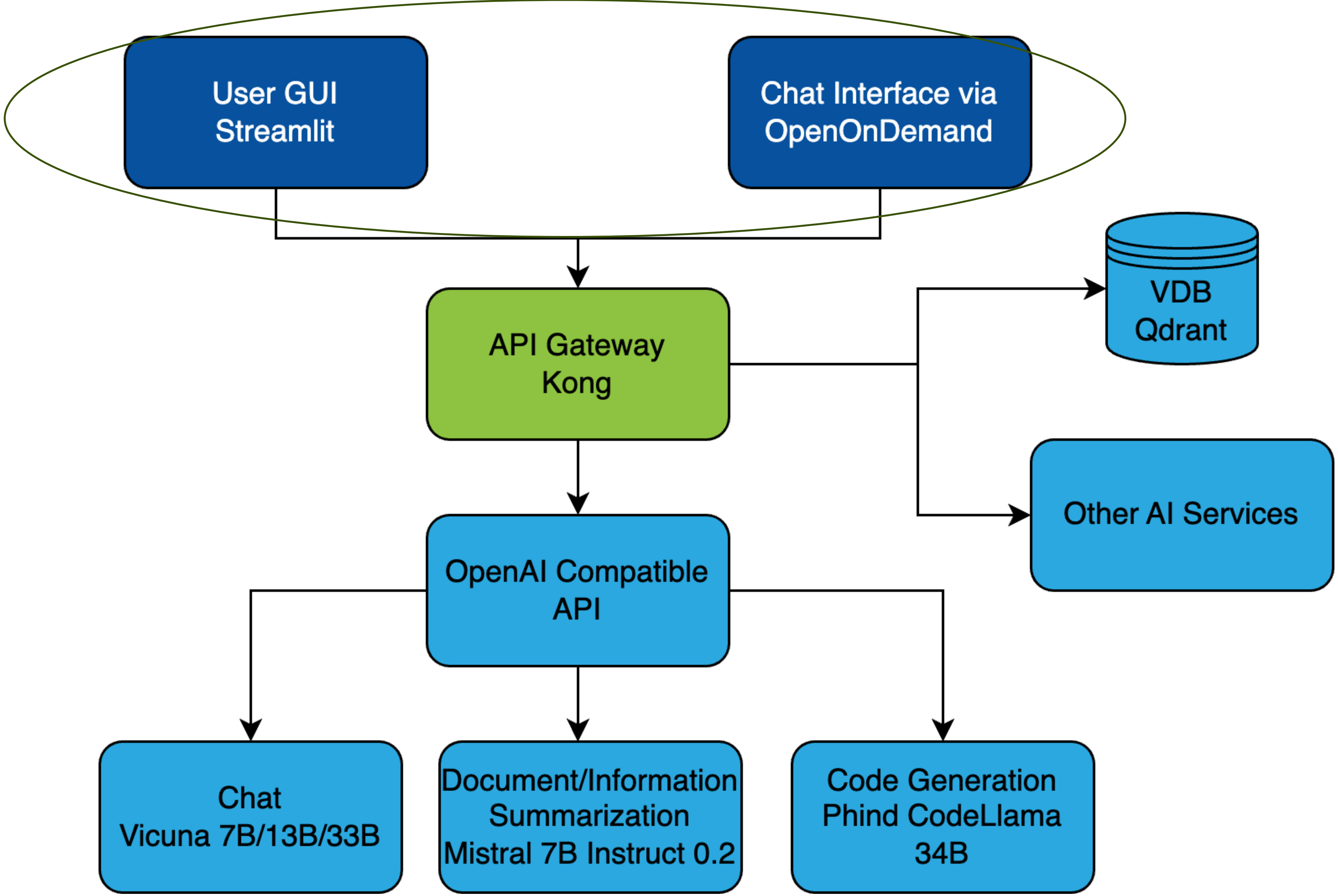
Overview

- On Premise vs Cloud
- Software Infrastructure
- Hardware
- Challenges
- Stats
- Questions

On Premise vs Cloud

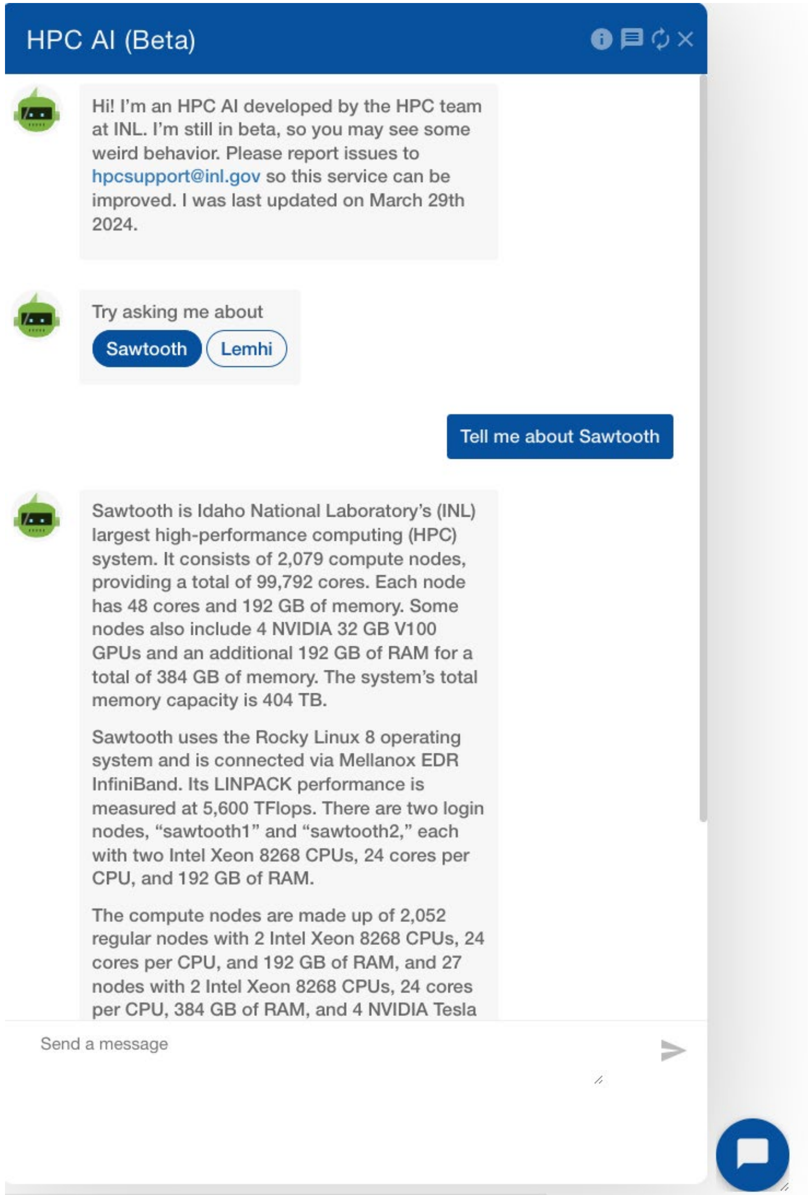
- Privacy was our #1 priority
- Partnerships with gov approved cloud providers take a substantial amount of time
- Costs
  - We already had purchased the hardware as it’s used for AI/ML research
  - Didn’t want to have (and don’t have the budget for) variable cloud API costs
- Using proprietary cloud models would probably give us better results (
  - Proprietary models are better than open-source models according to most benchmarks but the above were more important for us

User Interface



User Interface

- Streamlit for general chat
  - Very good for easy things, but VERY limiting after
  - https://streamlit.io
- Custom interface for Open OnDemand chat
- VSCode Extension
  - https://continue.dev



Hi! I'm an HPC AI developed by the HPC team at INL. I'm still in beta, so you may see some weird behavior. Please report issues to hpcsupport@inl.gov so this service can be improved. I was last updated on March 29th 2024.

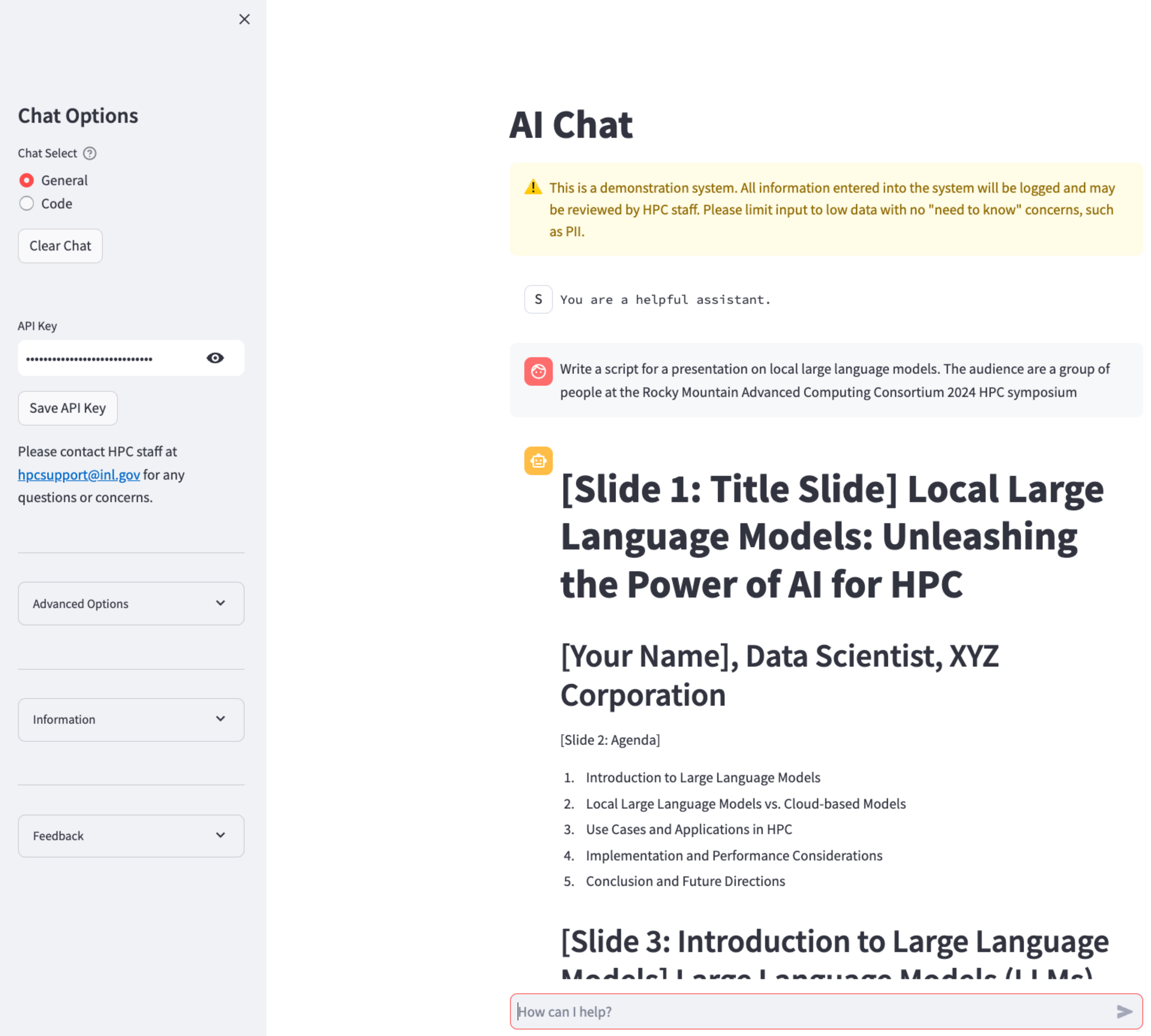
Try asking me about

Tell me about Sawtooth

Sawtooth is Idaho National Laboratory's (INL) largest high-performance computing (HPC) system. It consists of 2,079 compute nodes, providing a total of 99,792 cores. Each node has 48 cores and 192 GB of memory. Some nodes also include 4 NVIDIA 32 GB V100 GPUs and an additional 192 GB of RAM for a total of 384 GB of memory. The system's total memory capacity is 404 TB.

Sawtooth uses the Rocky Linux 8 operating system and is connected via Mellanox EDR InfiniBand. Its LINPACK performance is measured at 5,600 TFlops. There are two login nodes, "sawtooth1" and "sawtooth2," each with two Intel Xeon 8268 CPUs, 24 cores per CPU, and 192 GB of RAM.

The compute nodes are made up of 2,052 regular nodes with 2 Intel Xeon 8268 CPUs, 24 cores per CPU, and 192 GB of RAM, and 27 nodes with 2 Intel Xeon 8268 CPUs, 24 cores per CPU, 384 GB of RAM, and 4 NVIDIA Tesla



Chat Select

● General
○ Code

API Key: ••••••••••••••••••••••••••••

Please contact HPC staff at hpcsupport@inl.gov for any questions or concerns.

AI Chat

⚠ This is a demonstration system. All information entered into the system will be logged and may be reviewed by HPC staff. Please limit input to low data with no "need to know" concerns, such as PII.

S `You are a helpful assistant.`

Write a script for a presentation on local large language models. The audience are a group of people at the Rocky Mountain Advanced Computing Consortium 2024 HPC symposium

[Slide 1: Title Slide] Local Large Language Models: Unleashing the Power of AI for HPC

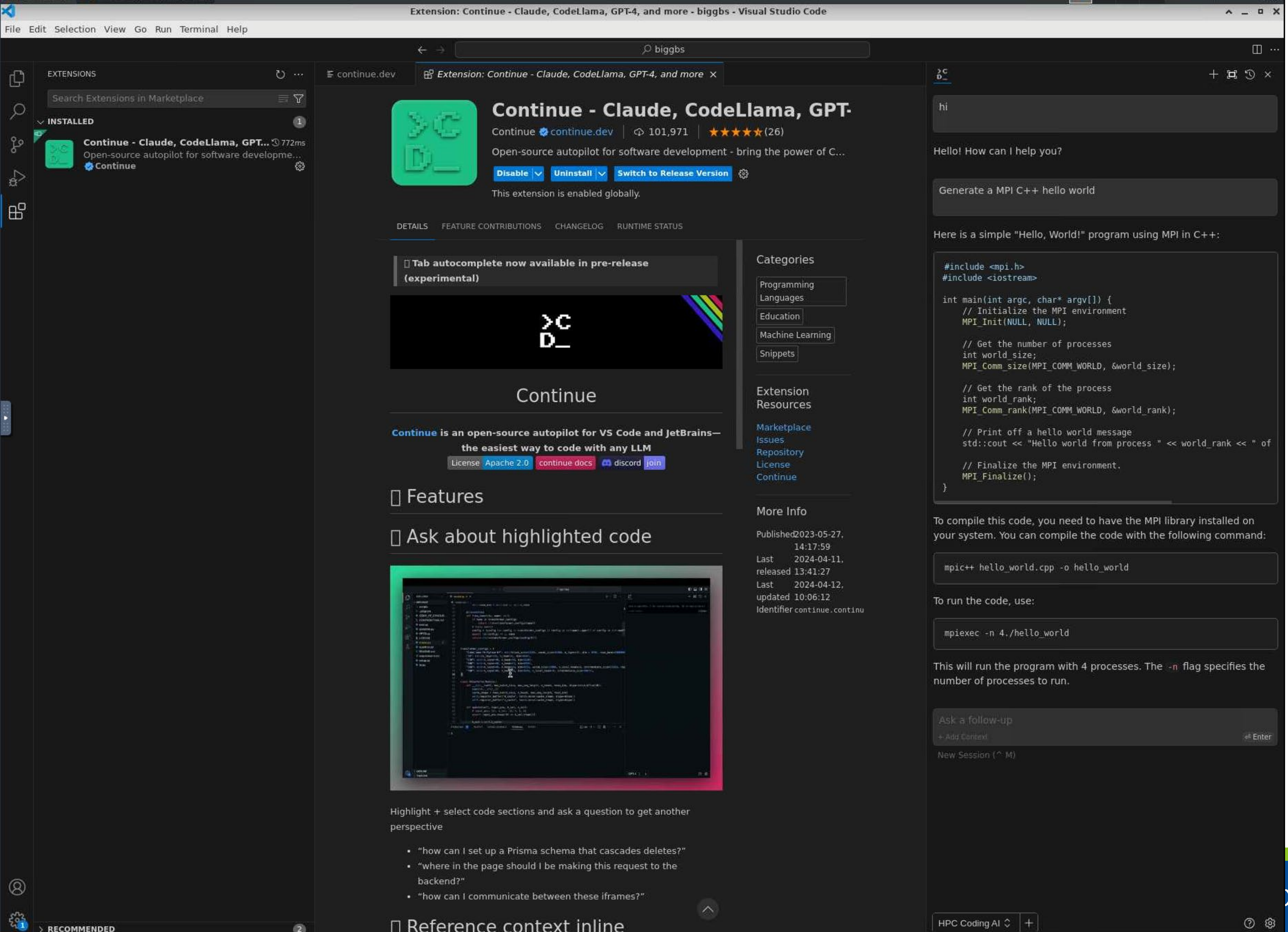
[Your Name], Data Scientist, XYZ Corporation

[Slide 2: Agenda]

1. Introduction to Large Language Models
2. Local Large Language Models vs. Cloud-based Models
3. Use Cases and Applications in HPC
4. Implementation and Performance Considerations
5. Conclusion and Future Directions

[Slide 3: Introduction to Large Language Models] Large Language Models (LLMs)

IDAHO NATIONAL LABORATORY


Extension: Continue - Claude, CodeLlama, GPT-4, and more - biggbs - Visual Studio Code
File
Edit
Selection
View
Go
Run
Terminal
Help
biggbs
EXTENSIONS
Search Extensions in Marketplace
INSTALLED
1
Continue - Claude, CodeLlama, GPT... 772ms
Open-source autopilot for software developme...
Continue
continue.dev
Extension: Continue - Claude, CodeLlama, GPT-4, and more
Continue - Claude, CodeLlama, GPT-
Continue
continue.dev
101,971
★★★★★(26)
Open-source autopilot for software development - bring the power of C...
Disable
Uninstall
Switch to Release Version
This extension is enabled globally.
DETAILS
FEATURE CONTRIBUTIONS
CHANGELOG
RUNTIME STATUS
Categories
Programming Languages
Education
Machine Learning
Snippets
Extension Resources
Marketplace
Issues
Repository
License
Continue
More Info
RECOMMENDED
2
Ask a follow-up
+ Add Context
Enter
New Session (^ M)
HPC Coding AI

□ **Tab autocomplete now available in pre-release (experimental)**

Continue

Continue is an open-source autopilot for VS Code and JetBrains—the easiest way to code with any LLM

License Apache 2.0 | continue docs | discord join

□ Features

□ Ask about highlighted code

Highlight + select code sections and ask a question to get another perspective

- "how can I set up a Prisma schema that cascades deletes?"
- "where in the page should I be making this request to the backend?"
- "how can I communicate between these iframes?"

□ Reference context inline

Published	2023-05-27, 14:17:59
Last released	2024-04-11, 13:41:27
Last updated	2024-04-12, 10:06:12
Identifier	continue.continu

hi

Hello! How can I help you?

Generate a MPI C++ hello world

Here is a simple "Hello, World!" program using MPI in C++:

```
#include <mpi.h>
#include <iostream>

int main(int argc, char* argv[]) {
    // Initialize the MPI environment
    MPI_Init(NULL, NULL);

    // Get the number of processes
    int world_size;
    MPI_Comm_size(MPI_COMM_WORLD, &world_size);

    // Get the rank of the process
    int world_rank;
    MPI_Comm_rank(MPI_COMM_WORLD, &world_rank);

    // Print off a hello world message
    std::cout << "Hello world from process " << world_rank << " of

    // Finalize the MPI environment.
    MPI_Finalize();
}
```

To compile this code, you need to have the MPI library installed on your system. You can compile the code with the following command:

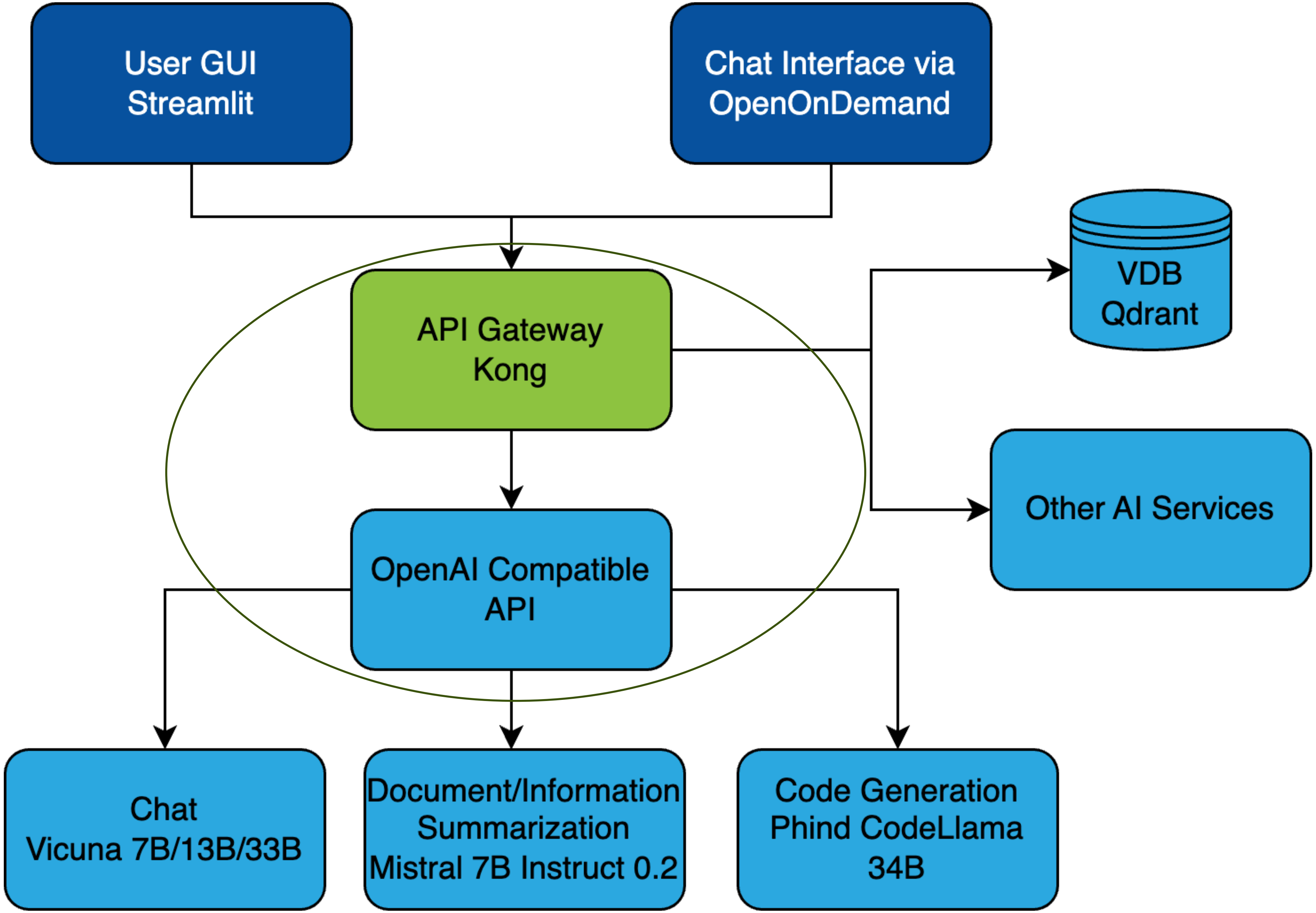
```
mpic++ hello_world.cpp -o hello_world
```

To run the code, use:

```
mpiexec -n 4./hello_world
```

This will run the program with 4 processes. The `-n` flag specifies the number of processes to run.

API



User GUI
Streamlit

Chat Interface via
OpenOnDemand

API Gateway
Kong

VDB
Qdrant

Other AI Services

OpenAI Compatible
API

Chat
Vicuna 7B/13B/33B

Document/Information
Summarization
Mistral 7B Instruct 0.2

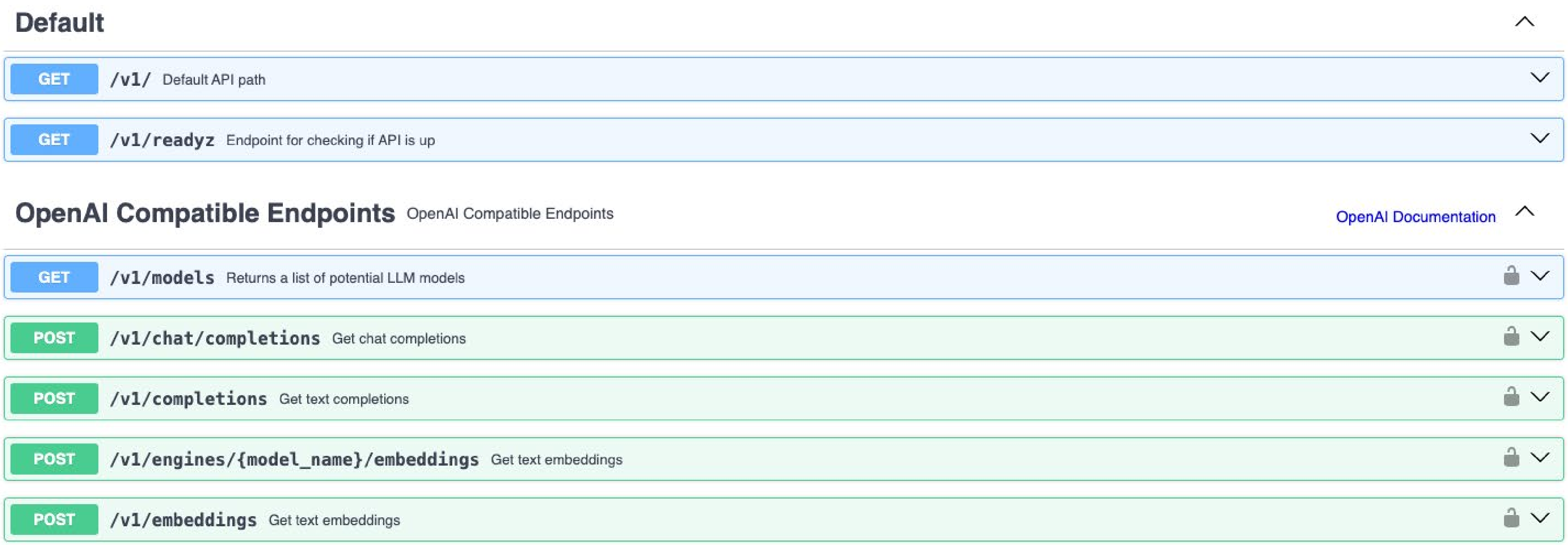
Code Generation
Phind CodeLlama
34B

API Tools

- Kong is an API gateway that sits at the front of all of our services
  - https://konghq.com/products/kong-gateway
  - Kong helps allows us to use services on a bunch of physical servers
  - Load balancing, rate limiting, and many additional plugins
  - Open-source version works great
- OpenAI compatible LLM API using lmsys Fastchat
  - https://github.com/lm-sys/FastChat
  - Provides multiple GPU support for bigger models
  - Can run vLLM under the hood
  - Has a built in user interface
  - Did not have API key support or great logging support, so we added it
  - Incredible tool overall though. Saved a ton of time and easy to use

API

- We give people API access and interactive documentation for the API
  - Allows people more flexibility in how they use LLMs
  - Have a few people using LLMs to categorize support tickets
  - Can use OpenAI API packages to access local API



Default

GET `/v1/` Default API path

GET `/v1/readyz` Endpoint for checking if API is up

OpenAI Compatible Endpoints OpenAI Compatible Endpoints — OpenAI Documentation

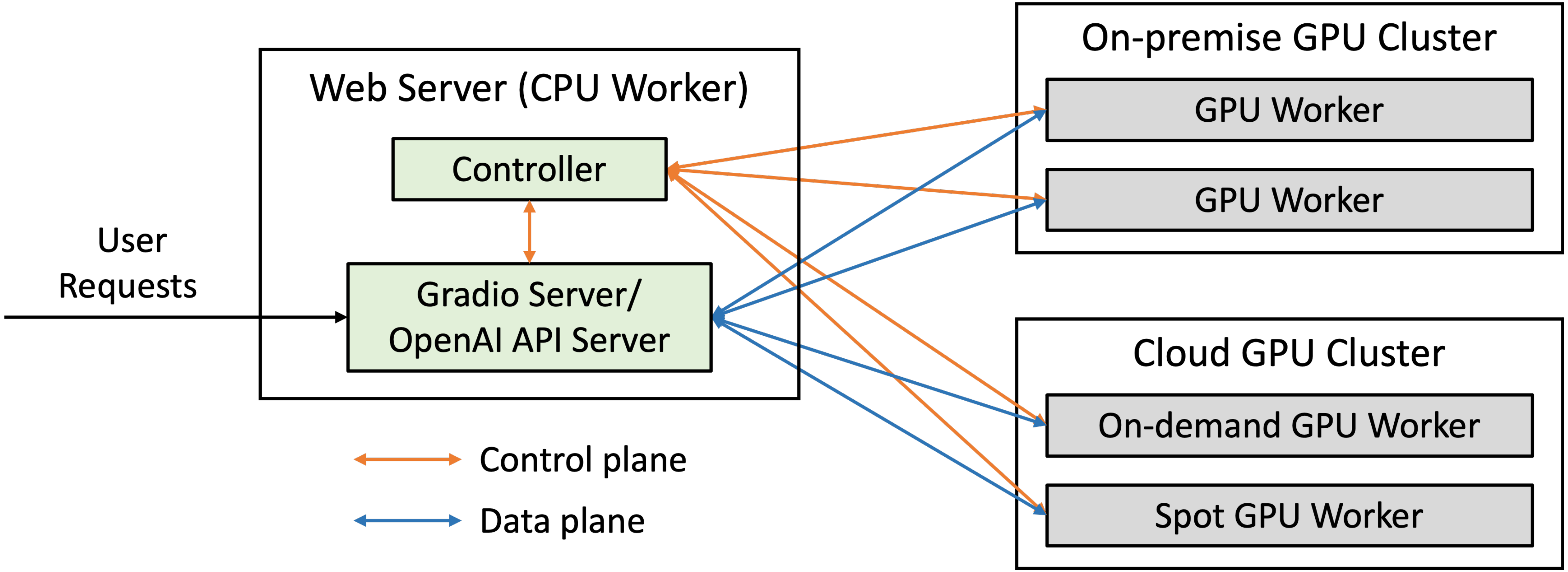
GET `/v1/models` Returns a list of potential LLM models

POST `/v1/chat/completions` Get chat completions

POST `/v1/completions` Get text completions

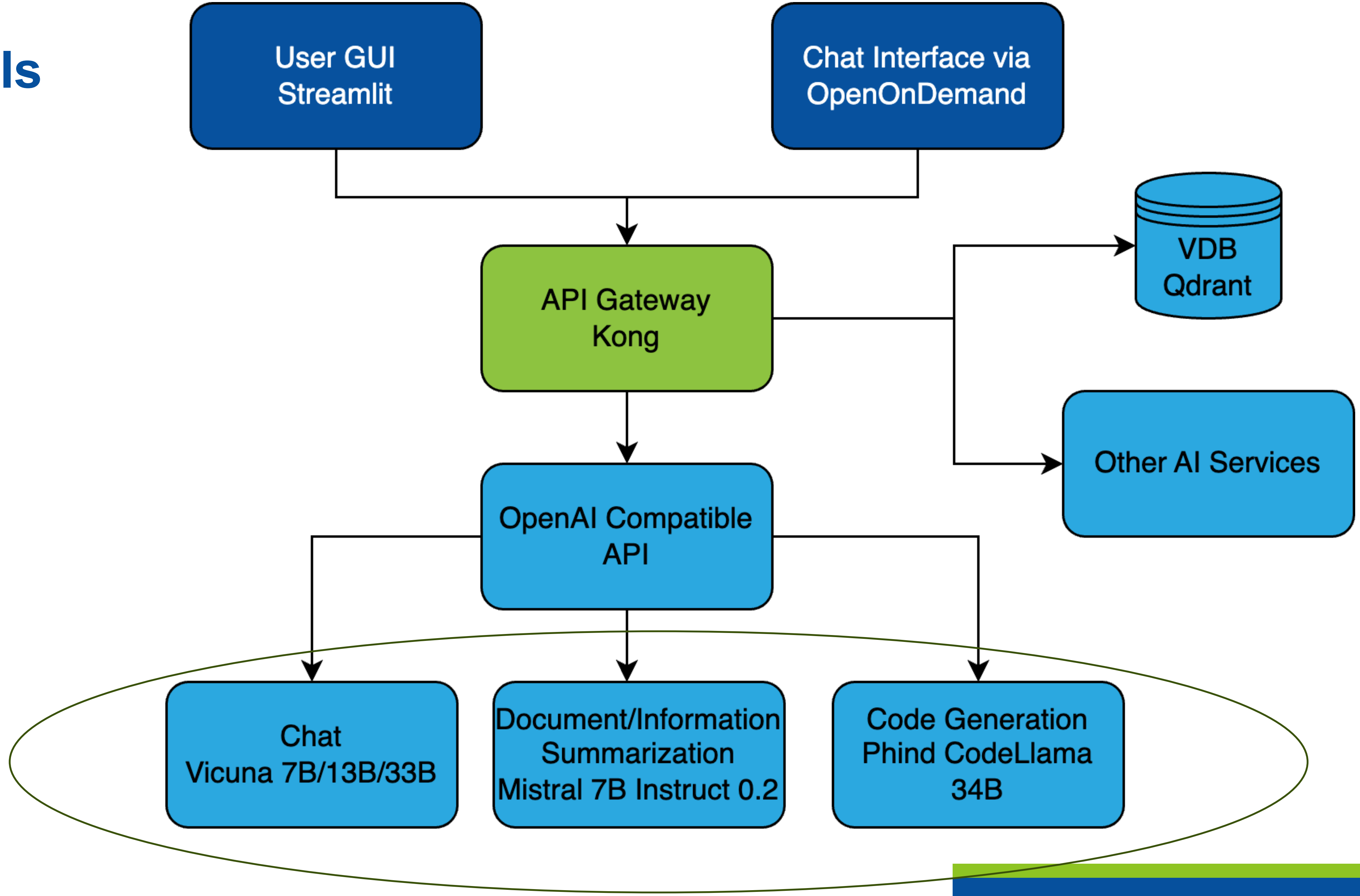
POST `/v1/engines/{model_name}/embeddings` Get text embeddings

POST `/v1/embeddings` Get text embeddings



https://github.com/lm-sys/FastChat/blob/main/docs/server_arch.md

Models


User GUI
Streamlit
Chat Interface via
OpenOnDemand
API Gateway
Kong
VDB
Qdrant
Other AI Services
OpenAI Compatible
API
Chat
Vicuna 7B/13B/33B
Document/Information
Summarization
Mistral 7B Instruct 0.2
Code Generation
Phind CodeLlama
34B

Models

- Vicuna from lmsys
  - 7B, 13B, 34B parameter models
  - https://huggingface.co/lmsys
  - Short context length but pretty good at following directions
- Phind-CodeLlama-34B-v2
  - https://huggingface.co/Phind/Phind-CodeLlama-34B-v2
  - Great for code documentation, quick scripts, bash, etc
- Mistral-7B-Instruct-v0.2
  - https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
  - 32K context window

Hardware

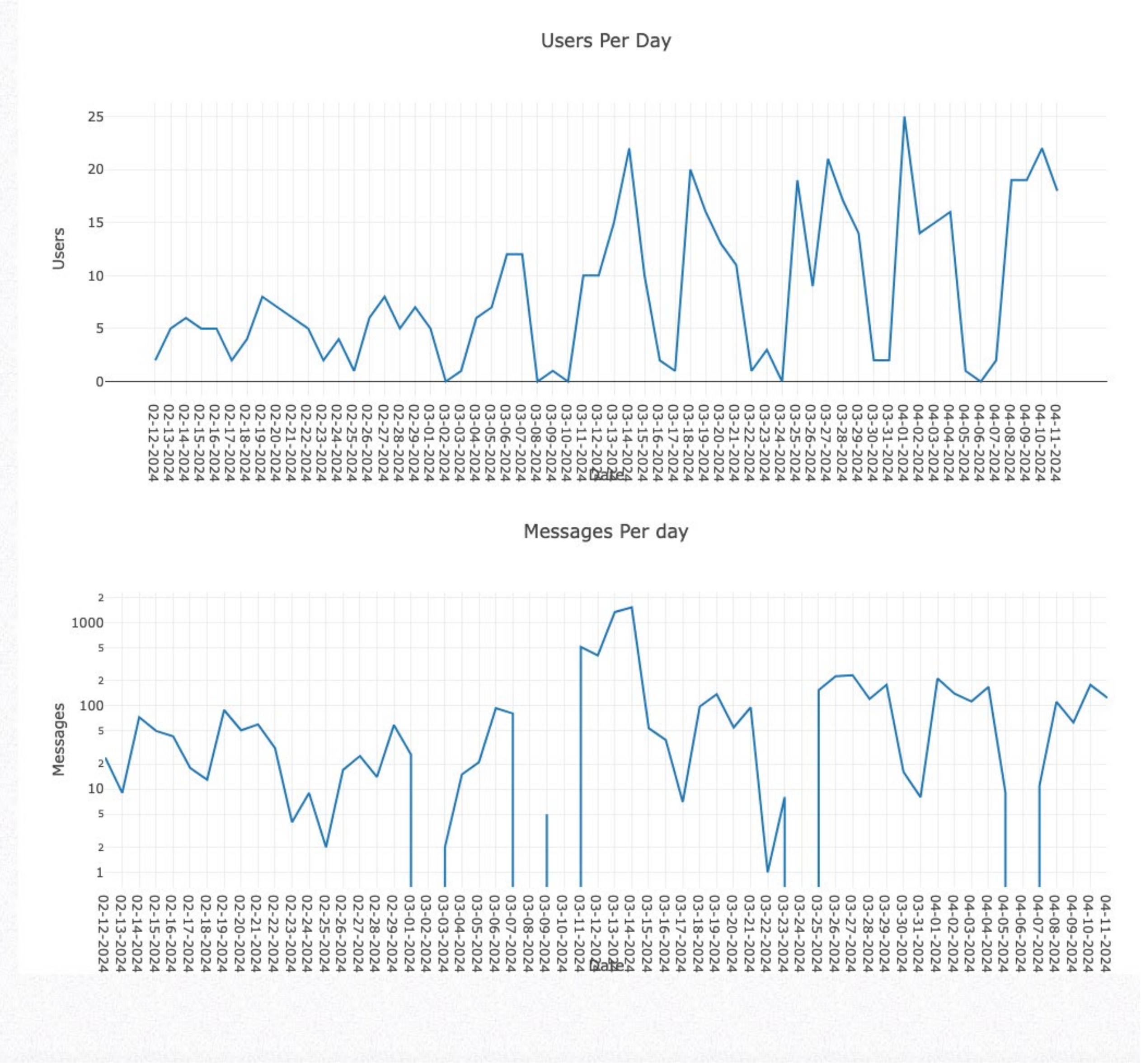
- Originally running on Nvidia V100s
- Moved to A100s
- Now running on H100s
- Haven’t tried AMD yet, but have an MI250 box on the way
- Haven’t ran any cross-node models yet

Challenges

- Security
  - API Keys which aren't perfect, but they help us keep track of people
  - This also means we must generate a key for whoever wants one. Haven't created a tool for self creation yet so doing it by hand
- Logging of messages took some time, especially for streaming messages
- Infrastructure
  - Not everyone has compute resources readily available
- We did have some issues initially with running out of GPU memory on 16GB V100s.
  - Could have been helped with quantized models (less precision) but lose some model performance
  - Moving to bigger/better GPUs solved the problem for us and we haven't had issues since (not yet at least)

Stats on Usage

- As of April 12, 11:04am:
  - Users: 86 (1.5% of total employees)
  - Messages: 7196
  - Tokens: 1,840,159
  - GPT4-32K pricing: $125.53
  - GPT4-Turbo pricing: $23.44
- If half of the lab was using it at the same rate:
  - GPT4-32K pricing: ~$4,000
  - GPT4-Turbo pricing: ~$1,000



Questions?

- Brandon.Biggs@inl.gov

Idaho National Laboratory

Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy. INL is the nation's center for nuclear energy research and development, and also performs research in each of DOE's strategic goal areas: energy, national security, science and the environment.