Grounded, Governed, and Scalable: LC's Approach to AI

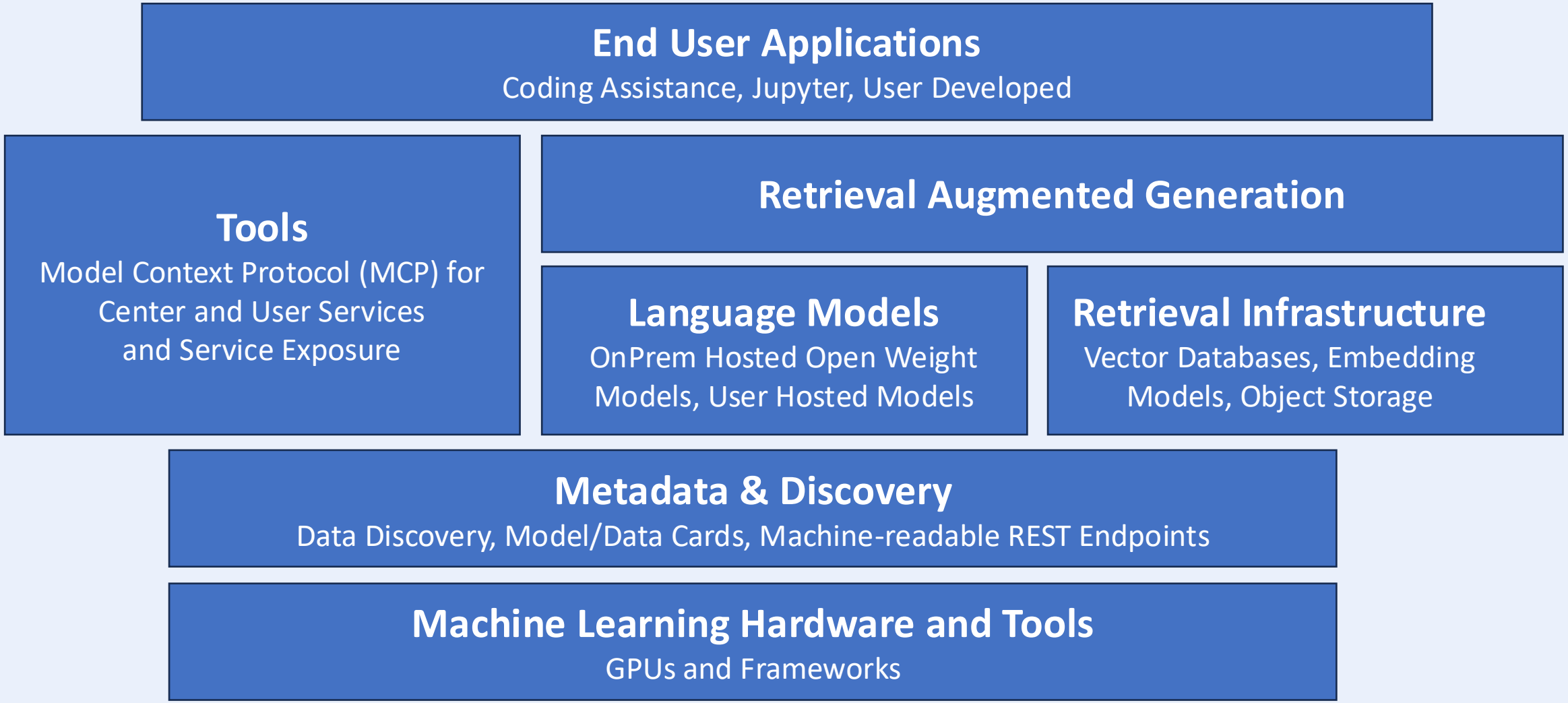
Computing External Review

April 2026

John Westlund
Livermore Computing, Workload Enablement Group

Prepared by LLNL under Contract DE-AC52-07NA27344.

Livermore Computing AI Ecosystem Overview



Hardware and Frameworks

I need hardware and software to train or fine-tune AI models on unique lab data

GPU/APU Accelerated AI Capable HPC Clusters

AMD

	Zone	Nodes	Total Memory (GB)	APU Architecture	GPU Architecture	Total GPUs
El Capitan	SCF	11,520	5,701,632	AMD MI300A	CDNA 3 [APU: AMD MI300A]	44,544
Matrix	CZ	30	8,064		NVIDIA H100	112
RZAdams	RZ	128	65,536	AMD MI300A	CDNA 3 [APU: AMD MI300A]	512

...

NVIDIA

AI Specialized Silicon

cerebras

sambanova

Prebuilt Accelerator Driver Containers and Environment Modules

NVIDIA CUDA

AMD ROCm

Prebuilt AI Frameworks

PyTorch

Metadata & Discovery

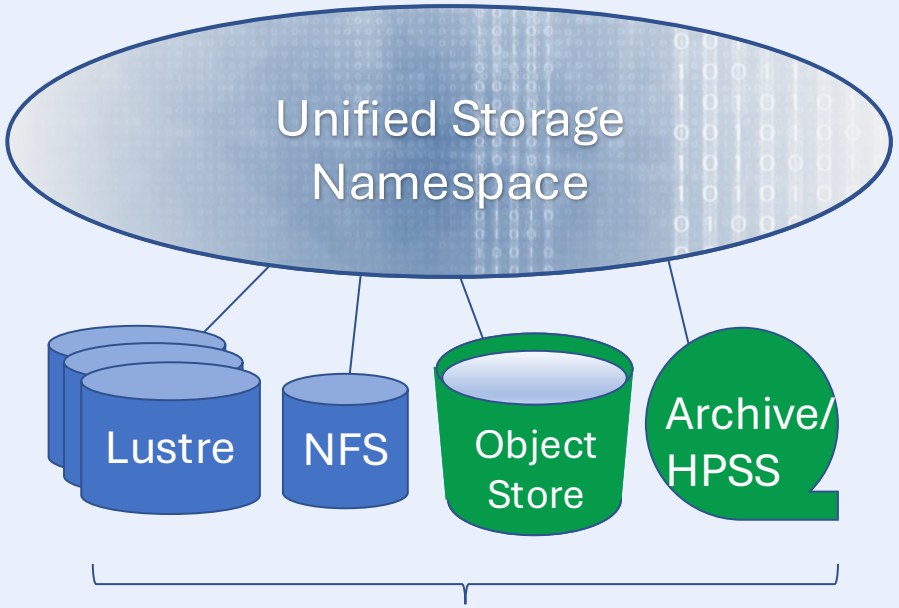
I need metadata services that my agents can query to find approved datasets and models

Model / Agent / Data Cards

```
language:
- en # ISO language tag
tags:
- project:genesis # include on all GENESIS project models
- project:model_team_name # include your _short_ model team
- type:model # use other types include {agent, eval, framew
- science:lightsource # what kind of science is this for (e
- risk:general # indicates level of risk review {general, r
license: {spdx_license_id} # use an SPDX license identifier
license_name: {license_name}  # If license = other (license
license_link: {license_link}  # If license = other, specify
base_model: {base_model} # if fine tuning, include the base
new_version: meta-llama/Llama-3.1-8B # if this model has be
datasets:
    - # a list of download URLs for dataset files used for
metrics:
    - # list of metrics used to monitor and evaluate model
    - # examples: training_loss, validation_accuracy, perpl
    - # specify metric sources or tools (e.g., WandB, Tenso
```

searchable and accessible via REST API

Unified Storage Namespace



USN Contains

- File/object names
- Source IOS
- Attributes (stat())
- Tags
- Key-value pairs

...

Metadata Only!

>1.5 EB capacity

O(10s) B files

Retrieval Infrastructure

I need a way to make my large datasets easily consumable by AI models

Retrieve Data

Amazon S3 Compatible Object Storage

VAST

HPSS

(S3 interface coming soon)

LC Network

LLNL Enterprise Network

Extract and Chunk Data

User facing service in planning:

(https://docling-project.github.io/docling/)

LaunchIT

Persistent Data Services

Choose the service you would like to deploy.

Service

LC LLamaMe API

Provided by LC IaaS (WEG)

API access to LC-hosted LLMs.

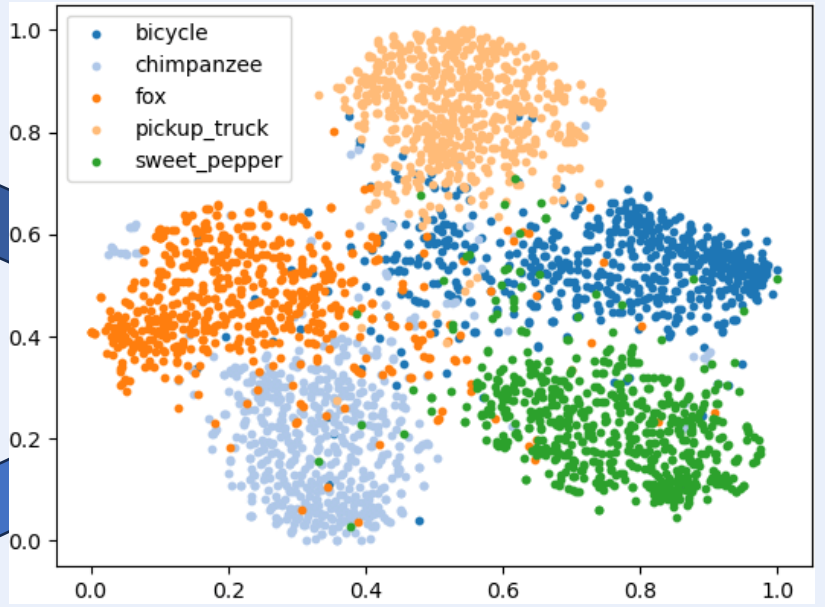
Service

LC PGVector PG-16 TLS

Provided by LC IaaS (WEG)

PostgreSQL database service with persistent storage and the postgres vector extension.

Generate and store embeddings



(Simard, Nathaniel & Lagrange, Guillaume. (2021). Improving Few-Shot Learning with Auxiliary Self-Supervised Pretext Tasks. 10.48550/arXiv.2101.09825.)

Language Models

I need to programmatically and securely interact with frontier, open-weight, and self-hosted models

1. LivIT

LLNL CUI APPROVED

LivAI API Model Listing

Comparison of LivAI API supported AI models

Updated: January 27, 2026

Lowest < $1 output | Low $1 - $3 output | Medium $3 - $10 output | High $10 - $20 output | Highest > $20 output

MODEL	PRICING (INPUT/OUTPUT)	CONTEXT	KNOWLEDGE CUTOFF	MODALITIES	FEATURES	COST CATEGORY
o1 OpenAI	$15.00 / $60.00	200K	Oct 01, 2023	TEXT, VISION	TOOLS, REASONING	HIGHEST
gpt-5.2 OpenAI	$1.75 / $14.00	272K	Aug 31, 2025	TEXT, VISION	TOOLS, REASONING	HIGH
gpt-5 OpenAI	$1.25 / $10.00	272K	Sep 30, 2024	TEXT, VISION	TOOLS, REASONING	HIGH

LivAI provided endpoints, industry leading FedRAMP High models, CUI only, $500/year provided to each user

2. hosted open weight model, "free" to LC users, CZ/RZ/SCF availability

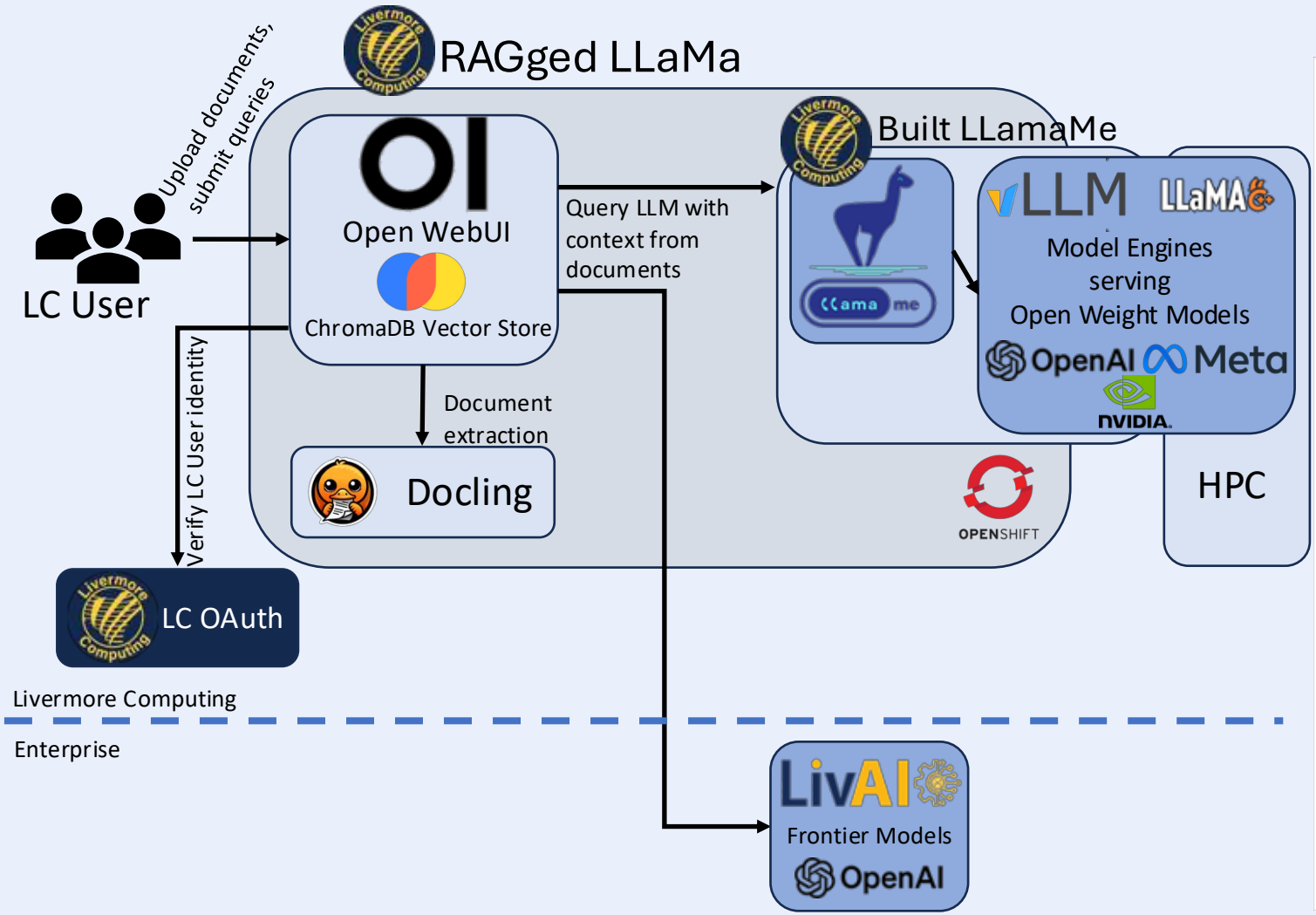
meta-llama/Meta-Llama-3.1-8B-Instruct (Max Model Length: 4096)	Default, OpenShift
Codestral-22B-v0.1 (Max Model Length: 32768)	RZVernal
gpt-oss-120b (Max Model Length: 131072)	RZVernal
gpt-oss-20b (Max Model Length: 128000)	RZVernal
intfloat/e5-mistral-7b-instruct (Max Model Length: 32768)	OpenShift
Llama-3.3-70B-Instruct (Max Model Length: 32768)	RZVernal
Llama-4-Scout-17B-16E-Instruct (Max Model Length: 128000)	RZVernal
nvidia/NVIDIA-Nemotron-Parse-v1.2 (Max Model Length: 9000)	OpenShift
_test-gpt-oss-120b (Max Model Length: 128000)	RZVernalTest

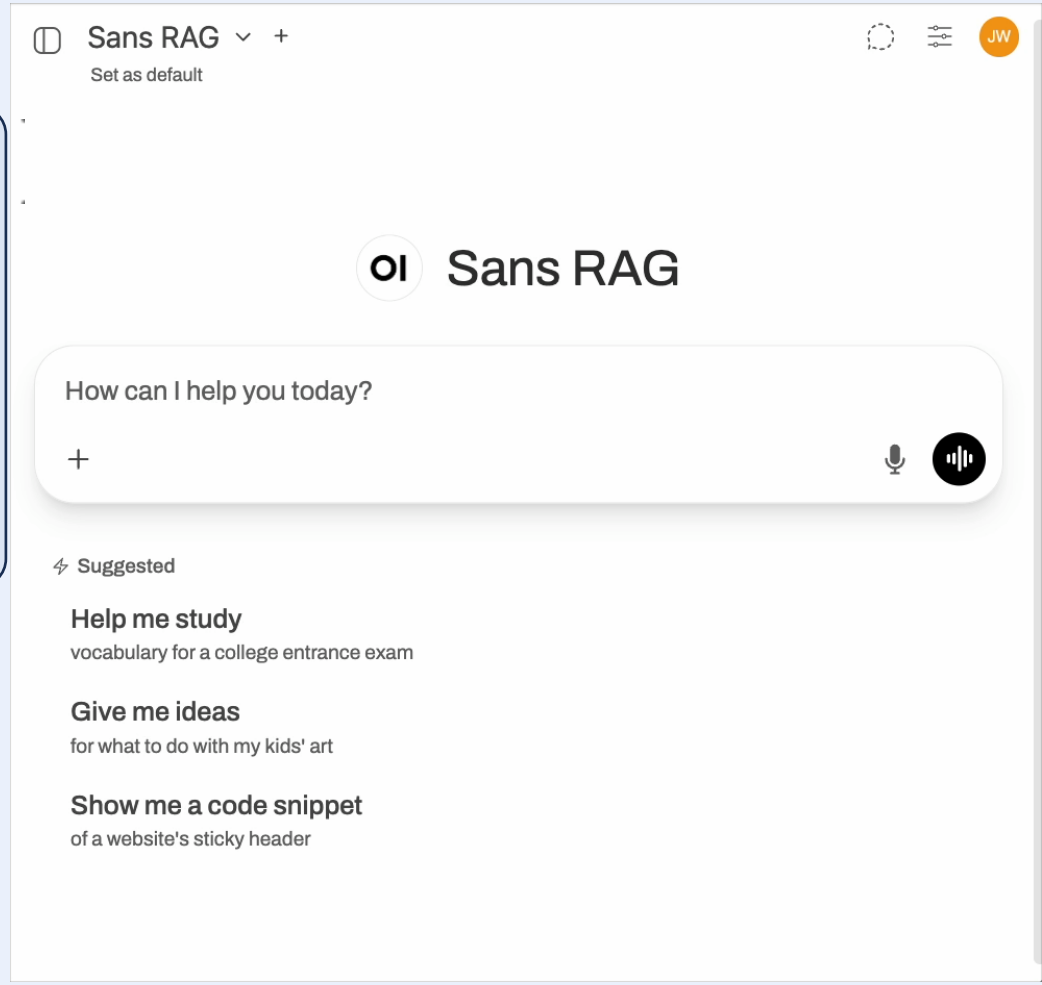
3. Mastodon

Facilitating users safely running models on hardware

Retrieval Augmented Generation

I need my own searchable knowledge workspace where I can upload domain-specific documents for my users





Limited availability for per project implementations

LC and User Provided Tools

I need a persistent environment where my services can run and communicate with HPC-provided system services

LC Wormhole

NNSA National Nuclear Security Administration
DAs / PAs

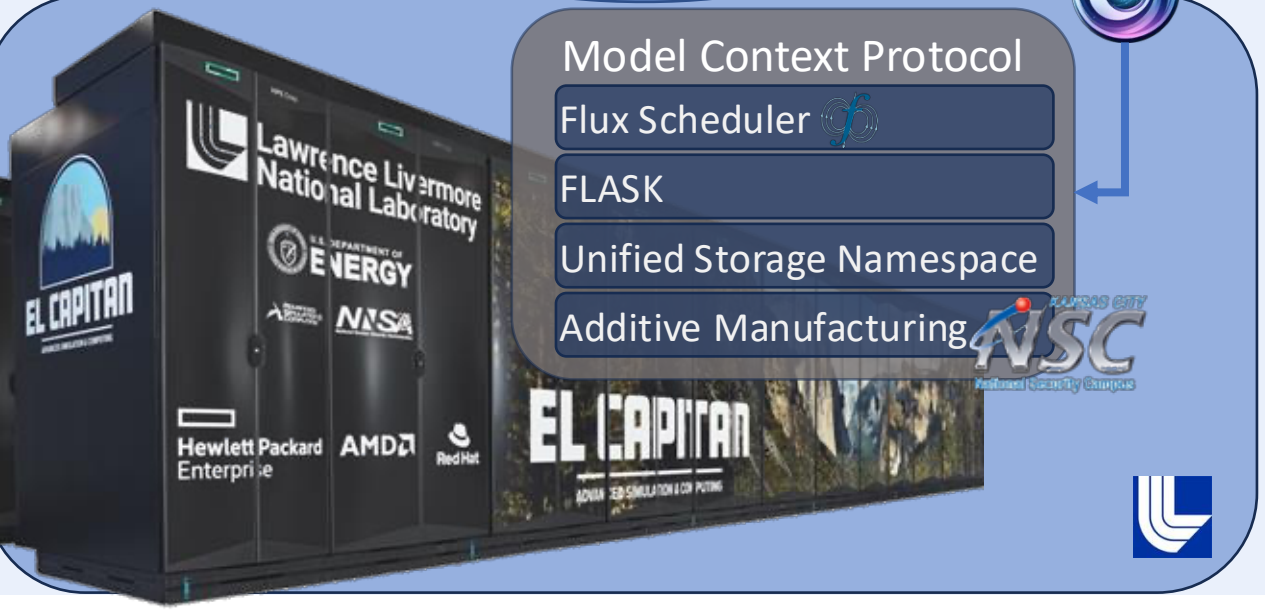
aws
LLMs, MCPs

oneID ESNhub One mission, infinite possibilities
OAuth2

Genesis Mission Agent Container

Persistent Data Services

Model Context Protocol
- Flux Scheduler
- FLASK
- Unified Storage Namespace
- Additive Manufacturing



OpenAI API

OpenAI o3

Model Context Protocol
- Slurm Scheduler

Wormhole High-Level

I need a uniform authentication mechanism so my services and users can securely access resources across network boundaries

Wormhole Speaks Enterprise

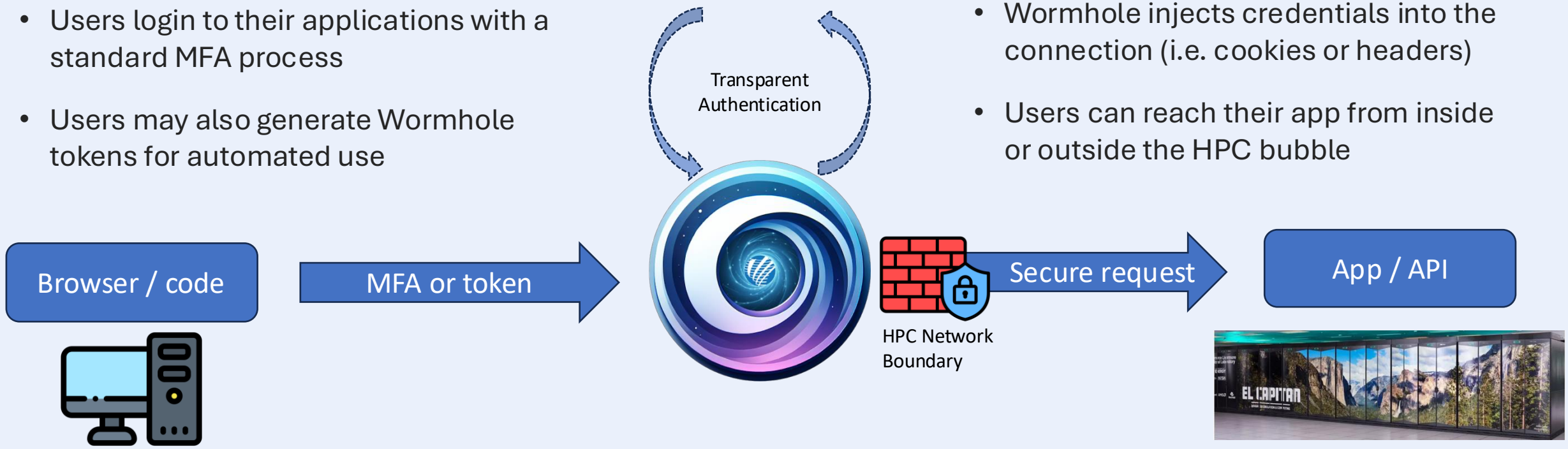
Wormhole bridges identity providers (OneID, ESN Hub ePKI, or site-specific) with apps and APIs

- Users login to their applications with a standard MFA process
- Users may also generate Wormhole tokens for automated use

Wormhole Speaks HPC

Wormhole automatically creates a pathway to the user's app, regardless of where it is running

- Wormhole injects credentials into the connection (i.e. cookies or headers)
- Users can reach their app from inside or outside the HPC bubble



End User Applications

I need an environment where I can get interactive assistance writing code and tools to visualize results

Codex

Sandboxed Coding Assistance

```
>_ OpenAI Codex (v0.46.0)

model:     gpt-5-codex high   /model to change
directory: ~/Desktop/igds-design-system

To get started, describe a task or try one of these commands:

/init - create an AGENTS.md file with instructions for Codex
/status - show current session configuration
/approvals - choose what Codex can do without approval
/model - choose what model and reasoning effort to use
/review - review any changes and find issues

› Use the Figma MCP Server to verify that the correct design variables are bein

• Updated Plan
  └ □ Inspect src/components/Button/Button.tsx to understand current design var
    □ Fetch design variable definitions from Figma for the button component.
    □ Compare implementation against Figma definitions and report discrepancies

• Explored
  └ Read Button.tsx, Button.css

• Calling figma.get_code_connect_map({"clientFrameworks":"react","clientLanguag

• Preparing to verify Figma design tokens (30s • esc to interrupt)

› Summarize recent commits
```

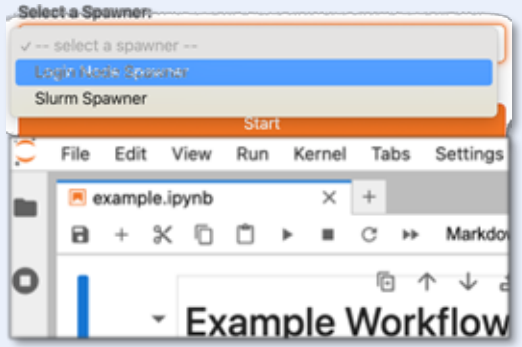
Models & Tools | Working Directories | Internet | Homedirectory

GitLab Duo

Repository integrated AI tooling

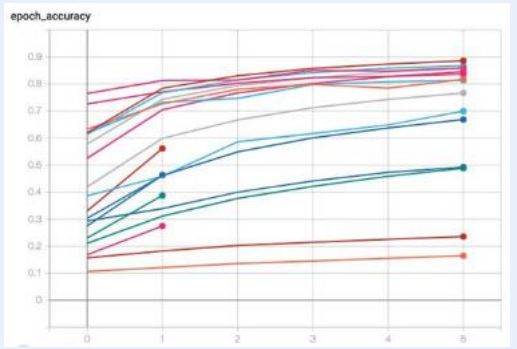
Orbit

Cluster1 ... ClusterN



HPC backed interactive coding environment

TensorBoard



AI training visualization

Workflow Enablement Group

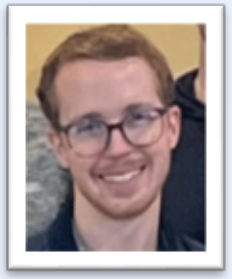
The nexus of cloud technologies, new workflows and traditional LC HPC ecosystems

WORKFLOW ENABLEMENT GROUP

A deep bench of diverse talent to aid HPC with challenges of disruptive technology

John Consolati (Team Lead)

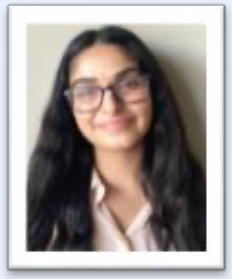
James Corbett



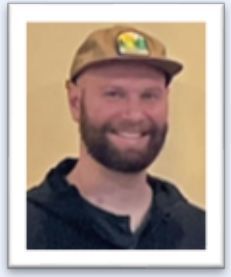
John Westlund

Todd Heer (Group Lead)

Urwah Mir

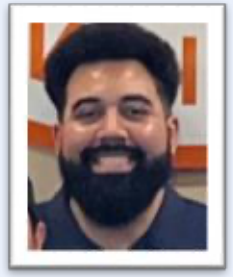


Zeke Morton



Conner Cook

Jordan Dorham

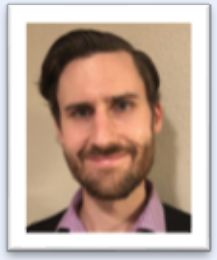


James Taliaferro

Otto Venezuela (Team Lead)

Crucial Cross-Group Collaborators/Contractors

Paul Bryant (ORNL)



Thomas Mendoza (contractor)

David Fox (Project Lead)