



LLamaMe (LLM) 3.0

LC User Meeting
October 6, 2026

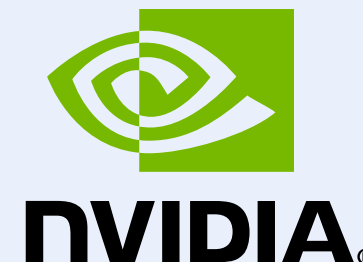
Kailey Wong, Otto Venezuela

LC Workflow Enablement Team | Platform Services Group

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

LLamaMe (LLM) API Service

- Providing users with access to locally-hosted open-source Large Language Models (LLMs) via REST API endpoint
- Allows users to integrate HPC jobs, integrated development environment (IDE), Retrieval-Augmented Generation (RAG) systems and LLM Agents
- Currently deployed using Nvidia GPU nodes in Kubernetes clusters and AMD nodes in EAS3 (Tioga, RZVernal, Tenaya) in the CZ, RZ, and SCF



General Availability in all 3 LC network zones (CZ, RZ, and SCF) since April 2025



Upcoming v3.0 Changes

01 — ENGINE

Custom JS proxy → LiteLLM

A well-maintained, industry-standard proxy replaces our in-house server.

02 — VISIBILITY

Usage & cost in LaunchIT

See token usage per key and model cost estimates, right where you manage keys.

03 — CAPACITY

Additional GPU Hardware

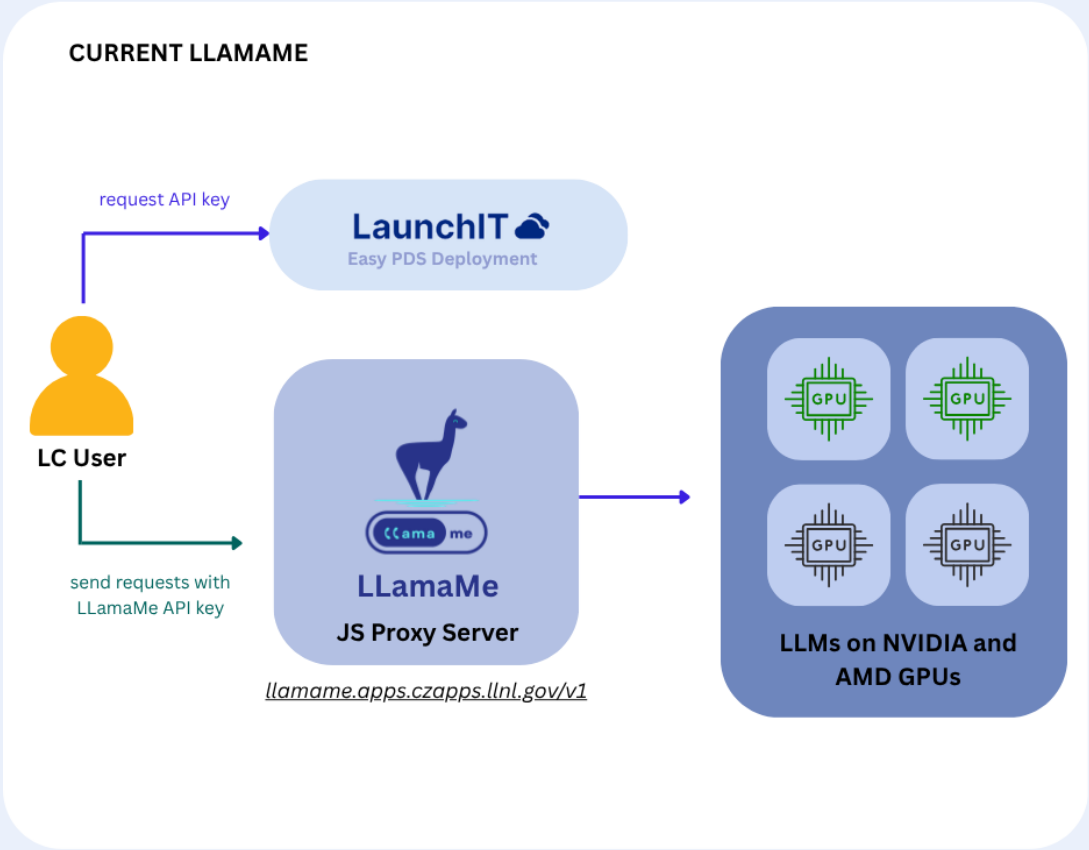
More hardware means increased capacity to support more (and bigger) models.

04 — NAMING

Model Renaming

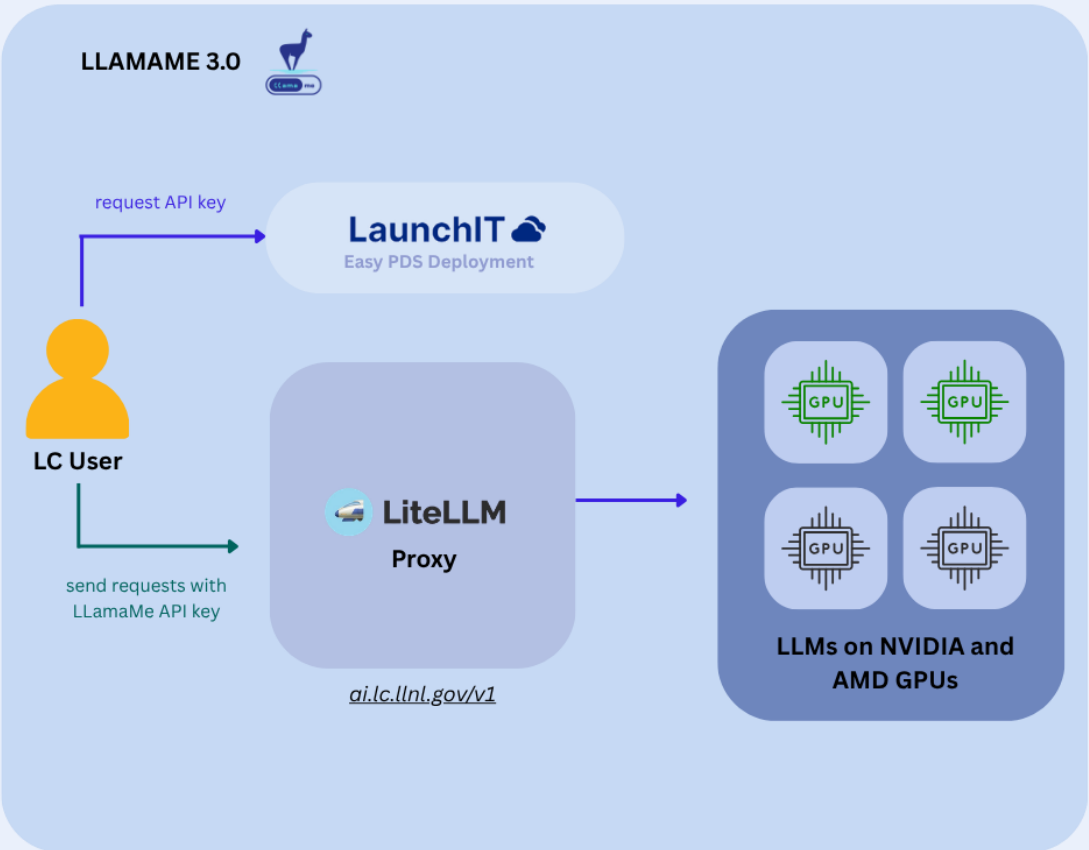
A consistent naming scheme across the all LLamaMe models and zones.

Custom JS Proxy → LiteLLM



llamame.apps.czapps.llnl.gov/v1

→



ai.lc.llnl.gov/v1

Same workflow, new endpoint!

Updated LaunchIT LLamaMe Dashboard

Resource Dashboard: LLamaMe API

[View Usage](#)[Back to cz-wong146 dashboard](#)

LLamaMe API:



LLamaMe API Key:

[REDACTED]



LLamaMe Endpoint URL:

https://ai.ic.llnl.gov/v1



LLamaMe Models:

gpt-oss-20b (Context: 128000)

Nemotron-Parse-v1.2 (Context: 9000)

Codestral-22B-v0.1 (Context: 32768)

gemma-4-31b-it (Context: 131072)

gpt-oss-120b (Context: 128000)

Llama-3.3-70B-Instruct (Context: 32768)

Low Cost

Openshift

Low Cost

Openshift

Medium Cost

TIOGA

Medium Cost

TIOGA

High Cost

TIOGA

Medium Cost

TIOGA

Generate new LLamaMe API key.

Your current LLamaMe API key will expire in 21 days on Oct 16, 2026. This action will generate a new LLamaMe API key for your workspace: **cz-wong146**

New LLamaMe API key

Delete this API key?

This action deletes your LLamaMe API key for workspace: **cz-wong146**

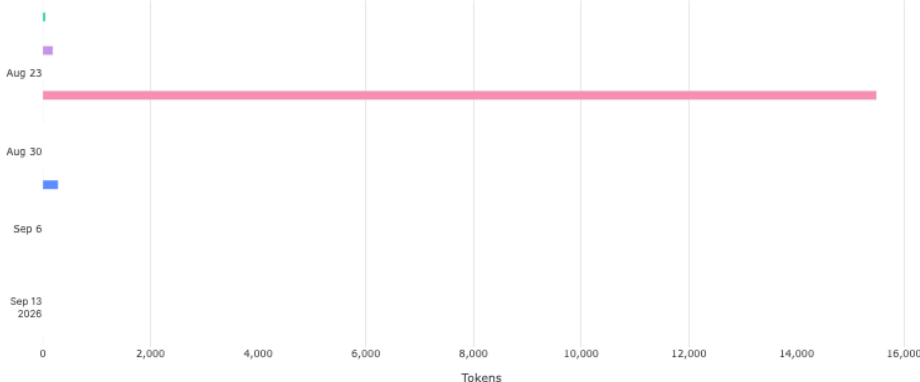
Delete

LaunchIT LLamaMe Usage Dashboard

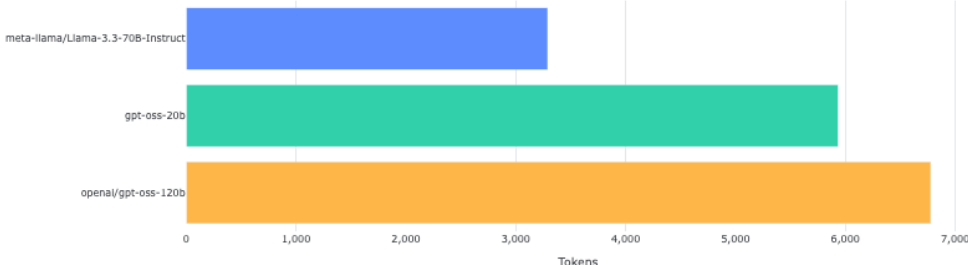
Usage Dashboard: cz-wong146 LLamaMe

[↩ Back to cz-wong146 LLamaMe dashboard](#)

Daily Token Usage



Tokens by Model



Model Usage Details

Model	Input tokens	Output tokens	Total tokens
openai/gpt-oss-120b	141	6,636	6,777
gpt-oss-20b	215	5,717	5,932
meta-llama/Llama-3.3-70B-Instruct	1,253	2,037	3,290

LaunchIT will shows token usage per day and per model

More Hardware!

- More headroom to help with load balancing
 - less queueing
- Room to bring on additional (larger) models
- Sits behind the same LiteLLM proxy
 - nothing additional you need to configure

Network	System	Added LLM Nodes
CZ	OpenShift	2x DELL H200
	Tioga	EAS3 - Mi250
RZ	RZVernal	EAS3 - Mi250
SCF	Tenaya	EAS3 - Mi250

NEW LLamaMe 3.0 Planned Models

INKLING

LAGUNA

MUSE
GLIMMER

NEMOTRON
3-Ultra

ColBERT

Your LLM suggestions welcome!

LLamaMe 3.0 Model Renaming/Decommissioning

Network	Model
CZ	meta-llama/Llama-3.3-70B-Instruct → Llama-3.3-70B Llama-4-Scout-17B-16E-Instruct → Llama-4-Scout-17B openai/gpt-oss-120b → gpt-oss-120B nvidia/NVIDIA-Nemotron-Parse-v1.2 → Nemotron-Parse-v1.2 nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4 → Nemotron-3-Super-120B gemma-4-31b-it → gemma-4-31b Codestral-22B-v0.1
RZ	meta-llama/Meta-Llama-3.1-8B-Instruct → Meta-Llama-3.1-8B meta-llama/Llama-3.3-70B-Instruct → Llama-3.3-70B Llama-4-Scout-17B-16E-Instruct → Llama-4-Scout-17B intfloat/e5-mistral-7b-instruct → e5-mistral-7b nvidia/NVIDIA-Nemotron-Parse-v1.2 → Nemotron-Parse-v1.2 gemma-4-31b-it gemma-4-31b Codestral-22B-v0.1
SCF	meta-llama/Llama-3.3-70B-Instruct → Llama-3.3-70B Llama-4-Scout-17B-16E-Instruct → Llama-4-Scout-17B intfloat/e5-mistral-7b-instruct → e5-mistral-7b gemma-4-31b-it gemma-4-31b Meta-Llama-3.1-8B-Instruct

Changes You'll Need to Make

- Create a new key in LaunchIT
 - no self-provisioning keys for shared workspaces/applications (for now)
- Update your endpoints
 - CZ: llamame.apps.czapps.llnl.gov/v1 → ai.lc.llnl.gov/v1
 - RZ: llamame.apps.rzapps.llnl.gov/v1 → ai.rzlc.llnl.gov/v1
 - SCF: llamame.apps.scfapps.llnl.gov/v1 → ai.lc.llnl.gov/v1
- Update any hard-coded model names

Stay connected ... call me (llámame)



Microsoft Team: [LC LLamaMe \(LLM\) | General | Microsoft Teams](#)



Docs: <https://hpc.llnl.gov/cloud>

GitLab: [LLamaMe Example Repo](#)



Feedback: lit-help@llnl.gov

Bugs: lc-hotline@llnl.gov

WE: otto@llnl.gov, wong146@llnl.gov

