



National Research Platform

7NRP WORKSHOP · MAY 5 TO 7, 2026

AI Inference and Infrastructure.

A tour of the open, federated LLM platform NRP runs for the research and education community · what it gives you and how it's built.

SPEAKER

Mohammad Firas Sada

AFFILIATION

UCSD / SDSC

SESSION

Thu May 7 · QI Auditorium

PART 01 • 15 MINUTES

The Stack.

An open Kubernetes cluster, federated across the world.

The **National Research Platform** is a single Kubernetes cluster that spans dozens of campuses on three continents. Every participating site contributes nodes · CPUs, GPUs, FPGAs, storage · to one shared API.

You don't see racks or data centers. You see `kubectl`. You write a manifest, and somewhere on the federation a Pod runs your workload · for free.

The same substrate that runs the LLM inference service runs everyone else's research workloads, side by side.

500+

NODES

CPUs, GPUs, FPGAs, storage ·
across all participating sites.

120+

LOCATIONS

Universities, national labs,
regional networks.

1

CLUSTER API

One `kubeconfig` reaches all of
it.

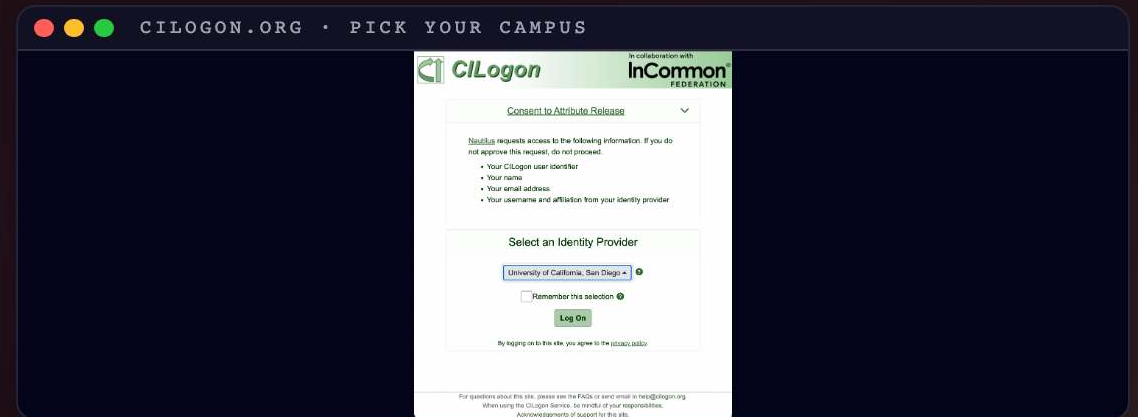
\$0

FOR ALL

Free at the point of use for
everyone.

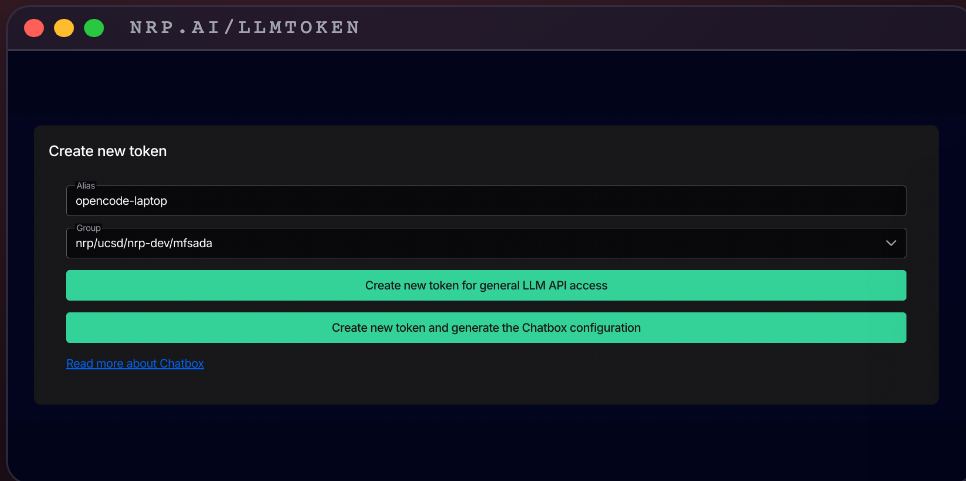
STEP 01 · SIGN IN

nrp.ai → Authentik → **CILogon** → your campus.



The **nrp.ai** Login button hands the OIDC flow to **Authentik** in the cluster, which delegates to **CILogon** · the InCommon-backed federation used by NSF/DOE. Pick your home institution from **200+ campus IdPs** and log in with your campus credentials. NRP receives an attested identity (email, affiliation, eduPersonPrincipalName) · **never your password**.

`nrp.ai/llmtoken` → OpenAI-shaped
curl.




WHAT'S HAPPENING

Pick the group, name the token, hit **Issue**. The token is **yours** · per-user, revocable, scoped to the group's flag.

```
curl https://ellm.nrp-nautilus.io/v1/chat/completions \
-H "Authorization: Bearer $NRP_TOKEN" \
-d '{ "model": "gwen3", "messages": [{ "role": "user", "content": "..."} ] }'
```

1,532 GPUs total · 74 reserved for LLMs · ~4.8%.

 NVIDIA · SUM OF nvidia.com/ ACROSS 248 GPU NODES · LLM COUNT = SUM OF STATEFULSET REPLICAS × PER-POD GPUS



GPU · Pod · Envoy · Auth · World.













Translated for the cabin: the model itself is locked inside Kubernetes · the only thing that can reach it is Envoy. Envoy is the bouncer. It checks your token with Authentik, decides if you're allowed in, picks the right Pod, copies your prompt over, streams tokens back, and writes down everything you spent. **One door. One auth. One log.**

↑ INSIDE THE POD:

vLLM is the serving engine we run for **shared inference**. For **training and fine-tuning**, NRP ships the tools and docs to make it easy to grab dedicated GPUs in your own namespace · same cluster, your weights, your schedule.

Ten frontier open-weights models on call, today.

 QWEN  GEMMA  GPT-OSS MOONSHOT · MINIMAX · ZHIPU · ALLENAI

MODEL	PARAMS	CONTEXT	STATUS
 qwen3	397B	262K	LIVE
 qwen3-small · qwen3-27b	27B	262K	LIVE
 gpt-oss	120B	131K	LIVE
 gemma	31B	262K	LIVE
 minimax-m2	230B	204K	LIVE
 qwen3-embedding	8B	—	LIVE
 kimi	1T MoE	262K	EVAL
 glm-4.7	358B	202K	EVAL
 olmo	32B	64K	EVAL
 gemma-small	8B	131K	EVAL



Aside from the API, NRP also **hosts web services** on top of the same models · Open WebUI, LibreChat, AnythingLLM · so users without code can drive the fleet from a browser.

PART 02 • 5 MINUTES

Usage in Practice.

71 teams · 38 institutions · Qwen3 takes 1/3 of all tokens.

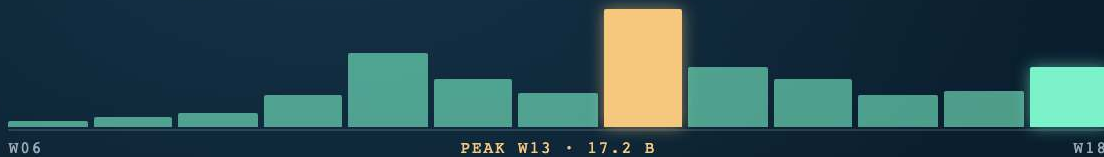
ADOPTION · 12 WK

14 → 71

TEAMS · W06 → W18

17.2 B

PEAK / WK · W13

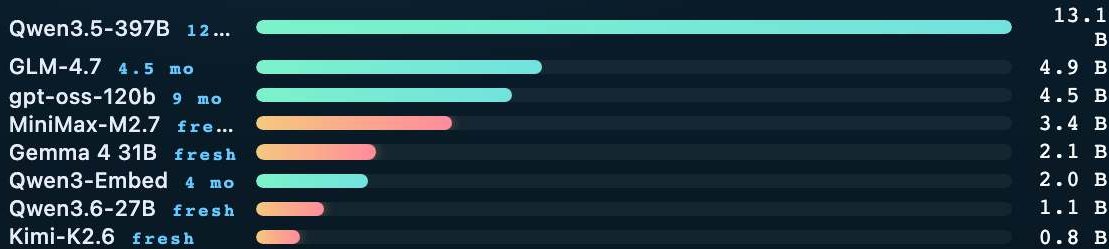


TOP INSTITUTIONS · 30D



+ 30 more · Princeton, Yonsei, Stanford, Internet2...

TOP MODELS · 30D



~36 B total · half < 6 wk old · SOTA in days

WHERE & SHAPE · 30D



95.5%

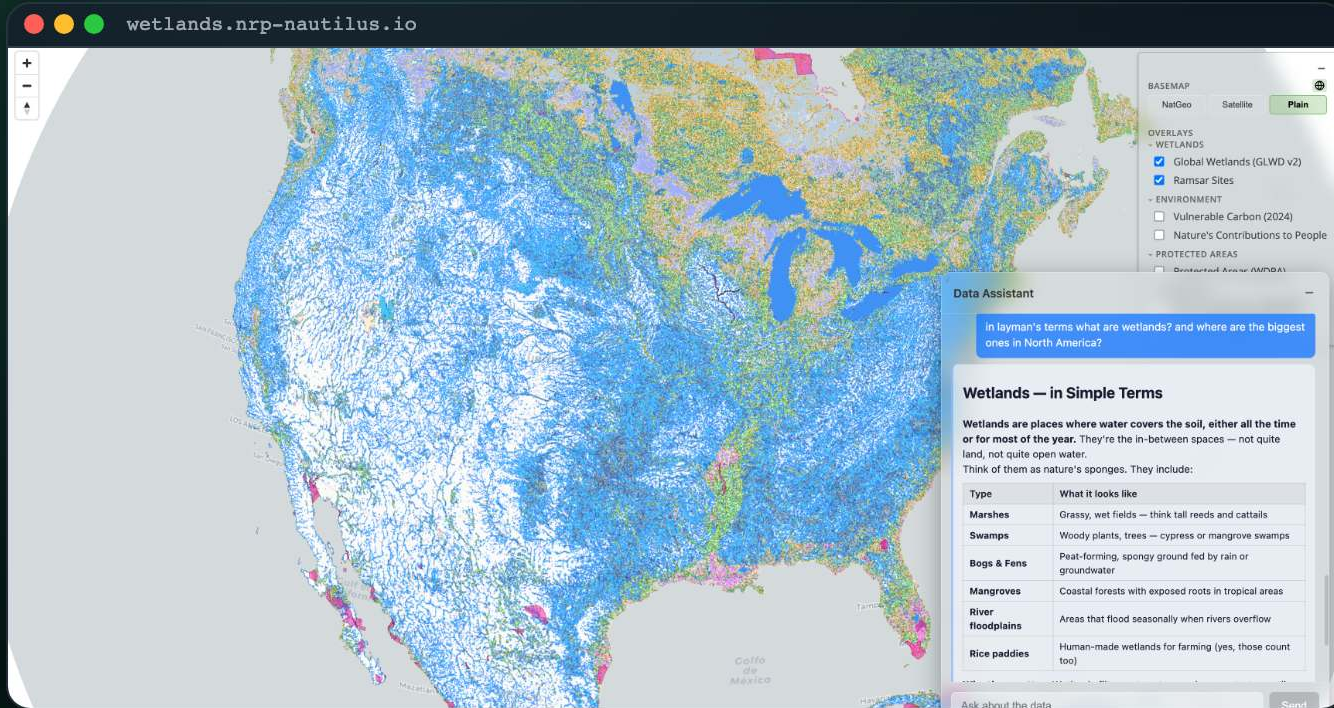
tokens are input

RAG & long-context · not chat-style

PART 03 • ~3 MINUTES

Compelling Use Cases.

Wetlands. One app, every NRP primitive.



Carl Boettiger

BOETTIGER LAB · UC BERKELEY · ESPM

WHAT IT IS

A user-facing app informing a global wetlands conservation effort · map + chatbot answers questions in plain language, grounded in real datasets.

STACK ON NRP

Static site · langchain-js chatbot on NRP LLMs · MCP server with [duckdb](#) over parquet on NRP S3 · STAC catalog · TiTiler for PMTiles/COGs · LLM proxy for rate limits.

NRP LLMs

MCP

S3 + STAC

TiTiler

duckdb

SIBLING APPS

[ca-wolves.nrp-nautilus.io](#)

[high-seas.nrp-nautilus.io](#)

Jupyter-AI plugin with Fernando Perez · NRP carbon dashboard.

The map displays the southern coast of California, from San Diego to Los Angeles. A blue line indicates the movement path of a wolf pack, starting near San Diego and heading inland. The map includes various geographical features like the San Pedro Channel, San Diego Trough, and San Clemente Basin. A 'Data Assistant' window is open on the right, providing information about the 'Yowlumni pack' and answering the question 'Should You Be Scared? No.' with four reasons. The window also includes a search bar and a 'Send' button.

Data Assistant

The **Yowlumni pack** (the southernmost) is around **448 km (~280 miles)** from La Jolla — and it's in the high Sierra Nevada mountains, not heading toward the coast.

🐾 Should You Be Scared? No.

Here's why:

1. **Wolves are extremely rare in California** — even in the core territory counties, confirmed sightings are rare events.
2. **Wolves actively avoid humans** — it's one of the main reasons wolf recovery is challenging. Attacks on humans are exceptionally rare across North America.
3. **La Jolla is not wolf habitat** — it's urban/suburban coastal San Diego. Wolves need large, remote wildland areas.
4. **Even in wolf country, your risk is essentially zero** — you are far more likely to be struck by lightning or attacked by a coyote than to encounter a wolf.

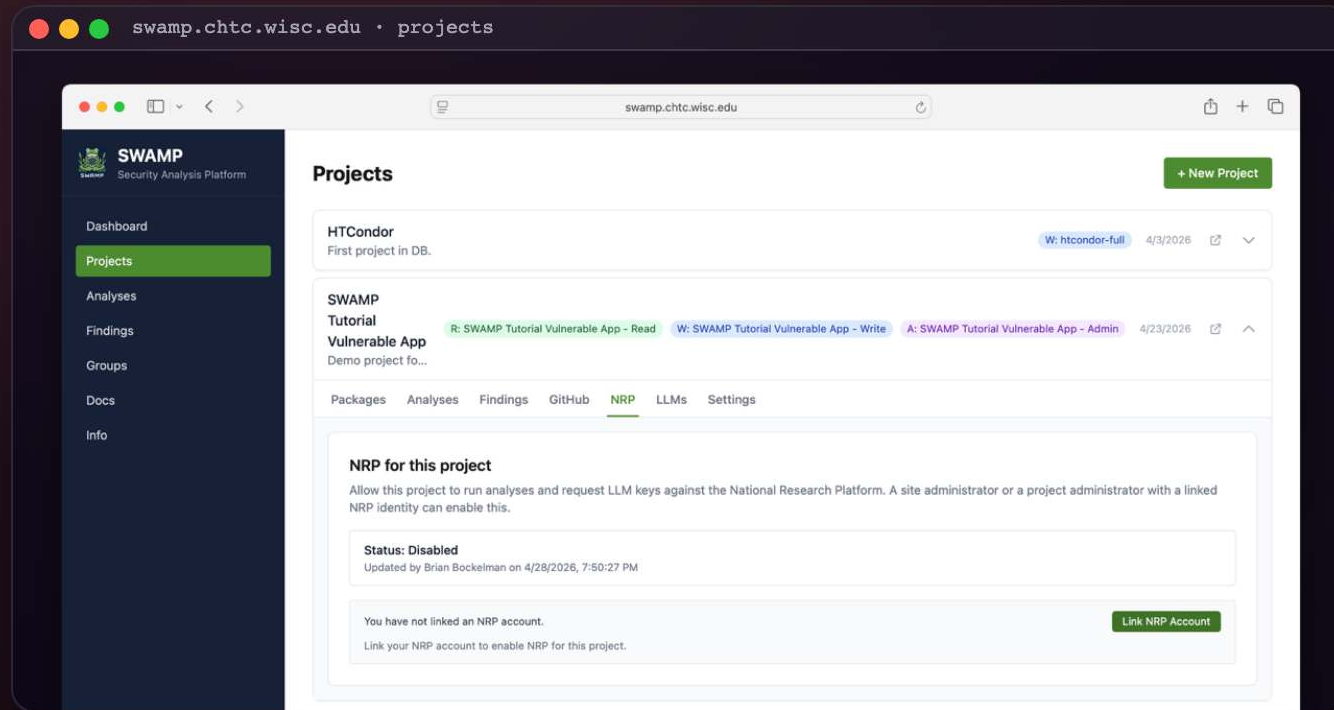
In short: Enjoy your beach day. The wolves are hundreds of

Ask about the data... Send

NRP MINIMAX M2.5

Wolf Movement (animated) [checkbox] Apr 21, 12:57 PM [1x]

SWAMP. Security analysis for any codebase · with NRP LLMs.



Brian Bockelman

CHTC · UW-MADISON · PELICAN PLATFORM PI

WHAT IT IS

An open-source platform for LLM-driven security reviews against **any** codebase. Vulnerability findings + concrete code-level remediation.

HOW IT PLUGS INTO NRP

Authenticate with your NRP identity · SWAMP issues a **project-scoped LLM key** bound to an NRP group · every analysis run uses NRP's frontier models, auditable, group-metered.

NRP identity

project-scoped keys

group accounting

WHERE IT LIVES

swamp.chtc.wisc.edu · run by the Center for High Throughput Computing at UW-Madison.

The screenshot shows a web browser window at `swamp.chtc.wisc.edu`. The page title is "SWAMP Tutorial Vulnerable App". The breadcrumb trail is "R: SWAMP Tutorial Vulnerable App - Read" > "W: SWAMP Tutorial Vulnerable App - Write" > "A: SWAMP Tutorial Vulnerable App - Admin". The date is 4/23/2026. The left sidebar has a menu with "Projects" (highlighted), "Analyses", "Findings", "Groups", "Docs", and "Info". At the bottom of the sidebar, it shows "Brian Bockelman (ORCID) user" and a "Sign Out" link. The main content area has a navigation bar with "Packages", "Analyses", "Findings", "GitHub", "NRP" (underlined), "LLMs", and "Settings". A green box at the top of the main content area says "Linked as `http://cilogon.org/serverA/users/9265706`". Below this is a section titled "NRP for this project" with the text: "Allow this project to run analyses and request LLM keys against the National Research Platform. A site administrator or a project administrator with a linked NRP identity can enable this." Underneath, it shows "Status: Enabled" and "Enabled by Brian Bockelman (ORCID) on 4/28/2026, 7:51:10 PM" with a "Disable NRP" button. A second green box shows a checkmark and "Linked as `http://cilogon.org/serverA/users/9265706`". The next section is "NRP LLM Key" with the text: "Exchange your linked NRP account for a project-scoped LLM API key. Once installed, the key is available to analyses on this project just like any other LLM provider." Below this, it shows "Currently installed: NRP LLM (osg-swamp) (...N24o)" and "Group `osg-swamp` will be used." with a green "Replace NRP LLM Key" button.

A node contributor also vibe-codes on the cluster.

```

opencode + kimi · debugging DeepBaseEditor
container: opencode — Konsole
New Tab Split View
seconds, 2.73 KB/s)
Now running DeepBaseEditor prediction... Please be patient...
File "/opt/DeepBaseEditor/wrapper.py", line 10
    print(f"Executing: {command}")
SyntaxError: invalid syntax

Thinking: We need to investigate errors in Python code under /home/ubuntu/persistent/container introduced by two specific commits. Also need to read config.yaml files under /home/ubuntu/persistent/deeppg/workflows and identify how each Python file is run and corresponding environment yaml files. We need to fix everything introduced by those two commits, preferably using a universally applicable solution to all files (such as changing wrappers to run on Python 2 to Python 3). Example error shows syntax error in /opt/DeepBaseEditor/wrapper.py because f-string not supported in Python 2, but the file runs under python2 maybe.

But the user says: "Read config.yaml files carefully under /home/ubuntu/persistent/deeppg/workflows and identify how each Python files here are run and the corresponding environment.yaml/environment-2.yaml/environment-3.yaml in /home/ubuntu/persistent/container/." So we need to understand workflow orchestration. Possibly Snakemake workflows or Cromwell? config.yaml may define steps that invoke python scripts, and environment yaml files may define conda envs for each step.

Need to scan everything. Let's first examine the directory structures, git history, and config yamls. Let's run some commands to list the files, check git log, see changes in those commits, and read the config files.

We have time to explore. We'll use bash to navigate.

# Show recent git log
$ cd /home/ubuntu/persistent/container && git log --oneline -20

35fba6f fix(model): CINDEL
1664d37 fix(model): Live output for logs
08e36b9 fix(ci): Add PYTHONUNBUFFERED
f182106 fix(ci): Make commit titles visible in CI
6310cda fix(model): Fail fast for DeepBE
e06c2e0 fix(model): Fail correctly for DeepPrime

Build · kimi NRP
~/persistent/container:main
    
```



Seungmin Kim · @ehfd

YONSEI GENOME EDITING LAB · PI: HYONGBUM HENRY KIM

WHO

Department of Pharmacology, Yonsei University College of Medicine. The lab contributes a node back to the federation.

HOW THEY USE NRP LLMS

Not for the science itself · for the tooling around the science.

Wrapper scripts, Python config debugging, conda environments, log triage. ~95% of that codebase is vibe-coded with kimi + opencode against the NRP gateway.

opencode

kimi

DeepBaseEditor

DeepPrime

WHY IT LANDS

Contributors are users. The same federation that runs your model shipped you the GPU it runs on.

Civilization V, played by LLMs.



WHO

John Chen · Assistant Professor, University of Arizona, College of Information Science.

WHAT IT IS

Civilization V opponents driven by LLMs over an MCP server.

Architecture: Civ5 ↔ Community Patch DLL ↔ Bridge ↔ MCP ↔ Vox Agents ↔ LLM. Built on Vox Populi.

THE RESEARCH QUESTION

Does ethical reasoning emerge when an LLM has to decide whether to launch a nuclear weapon · and which prompt interventions push it which way? Companion benchmark CivBench on arXiv.

NRP LLMs

MCP

CivBench

Civ V mod

WHERE IT LIVES

github.com/vox-deorum/vox-deorum · arxiv.org/abs/2604.07733

PART 04 · ~2 MINUTES

Agentic Coding · Live.

opencode + minimax-m2, building a flappy bird game at sunset.

PROMPT · SENT VERBATIM TO OPENCODE

NRP/MINIMAX-M2

```
› build a playable html5 flappy bird game with a sunset sky – bird, pipes, gravity, flap on space, score, game over +  
restart. one self-contained file.
```

▶ RUN IT LIVE · OPENCODE + MINIMAX-M2

🚣 Lifeboat · press me if you publicly embarrassed yourself by betting on a live demo not breaking

IDLE · TWO WAYS TO DEMO

OPENCODE · STREAMED VIA SSE

```
stdout / stderr will stream here when you press Run it live.
```

`nrp-accounting-mcp`, tools an agent can call.

01

Model Context Protocol

An open standard that lets language models call **typed tools** served by a remote process · same idea as a function-call API, but discoverable.

02

NRP runs one

`nrp-accounting-mcp.nrp-nautilus.io` exposes the cluster's accounting database. Tools like `top_resource_consumers` and `get_latest_data_date`.

03

Agent figures it out

You ask "who's using the most GPU lately?" · the agent picks the right tool, supplies the dates, formats the answer. No SQL, no schema lookup.

Full docs: nrp.ai/documentation/userdocs/ai/accounting-mcp/

Same agent, now talking to NRP itself.

PROMPT · SENT VERBATIM TO OPENCODE (WITH MCP TOOLS ENABLED)

NRP/MINIMAX-M2 + NRP-ACCOUNTING MCP

› who's using the most GPU on NRP lately? show me the top 5

▶ RUN IT LIVE · AGENT + MCP

🚫 Lifeboat · press me if you publicly embarrassed yourself by betting on a live demo not breaking, AGAIN

IDLE · AGENT IS ONE CLICK AWAY

OPENCODE · MCP TOOL CALLS + AGENT LOG

tool calls and agent log will stream here when you press **Run it live**.

Raw response from minimax-m2

▶ SEE RESULTS

RUN THE AGENT OR LOAD THE CACHE · THE RAW MODEL OUTPUT WILL APPEAR HERE

Questions?

- NRP managed LLMs · nrp.ai/documentation/userdocs/ai/llm-managed
- Token at · nrp.ai/llmtoken
- Usage at · grafana.nrp-nautilus.io

SPEAKER

Mohammad Firas Sada

EMAIL

mfsada@ucsd.edu