LLM proxy

One open-llm-proxy Deployment sends each chat request to a provider by model id, and logs every request and response.

ROUTING

GLEN apps

Each app's nginx forwards /api/llm/ and adds the key.

Benchmark runner

Headless Jobs that drive the same agent loop.

Other clients

Anything holding the proxy key.

OpenAI-style chat request

open-llm-proxy

DEPLOYMENT ×3

biodiversity namespace, one in-cluster service

1. Check the proxy key
2. Match the model id to a provider
3. Forward with that provider's key
4. Log the request and the response

`boettiger-lab/open-llm-proxy`

by model id

NRP hosted models

Open-weight models served on NRP. Any id not claimed by another provider goes here.

Commercial APIs

OpenRouter for ids with a vendor prefix, and Anthropic for claude ids.

`z-ai/` `deepseek/` `qwen/` `claude-`

DSE hardware

vLLM on the lab's own GPU servers, matched by exact id.

`nimbus` `cirrus`

LOGGING

EVERY 60 SECONDS

Raw JSONL

Each worker flushes its records to the log bucket on NRP Ceph, with credentials scrubbed.

`logs-open-llm-proxy`

DAILY AT 03:00 UTC

Daily Parquet

A CronJob rolls the day into tables of entries, of turns and of sessions.

`consolidate-daily`

MONTHLY ON DAY 2

Monthly Parquet

Completed months merge into one file each, kept indefinitely.

`consolidate-monthly`

HOURLY

Read-only mirror

Today's records copy to rustfs, where a scoped key can read them but change nothing.

`mirror-raw-hourly`