Benchmark and glen-in-a-box

The same parts serve two more jobs. A benchmark drives the real agent against the live services, and a single machine can run the whole stack offline.

EVALUATION

PRIVATE REPO WITH A PUBLIC DASHBOARD

glen-benchmark

Question bank

175 questions from 8 apps, each with a gold answer and runnable SQL.

Headless runner Jobs

The browser's agent code runs as a matrix Job, one cell per question, model and trial. One run had 688 cells.

`open-llm-proxy` `geo-query` `same STAC and S3`

Grading

66 questions grade mechanically against the gold, and 109 by a judge model.

Model leaderboard

Accuracy, consistency, completion, cost and latency per model. Transcripts stay private.

`boettiger-lab/glen-benchmark` `glen-benchmark-runs`

PORTABILITY

DOCKER COMPOSE ON ONE MACHINE

glen-in-a-box

Every part runs with the network cable unplugged, each standing in for its cluster counterpart.

one machine

webapp
nginx with the app and vendored JS, in place of the CDN

llama.cpp
in place of the proxy and models

geo-query
the same MCP image

rustfs
in place of NRP Ceph S3

TiTiler and basemap
local raster and base tiles

`field sites` `secure facilities` `classrooms` `conference demos`

`SchmidtDSE/glen-in-a-box`