Scientific environments

DATA SPECIFICATION / VERSION 0.2.0 / 17 OCTOBER 2026

Agents choose informative experiments, interpret noisy results, and predict unmeasured outcomes under a fixed budget. Three original synthetic mathematical testbeds, built with **Karotte 3.0.72**.

Environment	Hidden state	Credits	Submission
Causal circuits	7 coefficients; public nonlinear graph	8	8 x 2 values
Kinetic discrimination	Serial/parallel mechanism; 2 rates	3	15 x 1 values
Adaptive spectroscopy	9 candidate libraries; mixture/background	8	12 x 1 values

Episode and data contract

lab_state() supplies model/prior, action catalogue, free initial observations, target inputs, output order, noise, budget and history. **run_experiment(action_id)** returns query, values, noise_std, action_id, replicate, cost and budget_remaining. Each accepted assay costs one credit; maximum two executions per action. Invalid actions cost nothing and do not advance noise. Re-reading initial data adds no measurements.

Submission and reward

One terminal JSON file: `{"predictions": [[number, ...], ...]}`. Exact dimensions and finite real numbers are required. Reward = **1 / (1 + mean(((prediction - target) / scale)^2))**. Targets are noiseless means; scales are 0.035 / 0.020 / 0.025, respectively. Report reward, raw RMSE, normalized MSE and experiments used. Budget enforces cost; unused credits earn no bonus. Two causal output entries are publicly specified clamp values.

Generation and separation

Fresh protected seeds generate independent parameters and Gaussian measurement noise from published priors. Train/test draws share those priors; this is not out-of-family generalization. Target inputs are visible; episode parameters, seeds and target values remain teacher-owned. Public browser examples intentionally disclose fixed fixtures and are not protected evaluation episodes.

Delivery and acceptance

Each task includes its generator, Karotte tools/grader, public schema, replayable trace, reference policies, per-world results, tests, provenance and limitations. The 50-task catalogue contains **3 implemented prototypes + 47 proposed designs**; seeds do not count as distinct tasks. Agree the buyer's agent, difficulty band, scientific review, rights and acceptance tests before commissioning expansion.

Evidence boundary

Numerical, replay, budget and submission checks are documented in the release report. Statistical baselines and in-session model attempts are separate evidence. No physical calibration, RL training gain, external human expert endorsement, comprehensive security audit or buyer acceptance is claimed.

Research anchors: Foster et al. (ICML 2021); Tigas et al. (ICML 2023); Nabika et al. (Scientific Reports 2024). Full source mapping and 2023-2026 research: ohm.cx/projects/scientific-environments/evidence.