LLM EVALUATION PLATFORM

Eval

Rigorous, reproducible evaluation at scale. Test trajectories, responses, and model behavior without managing infrastructure.

Evaluation shouldn't require infrastructure

Running rigorous LLM evaluations at scale means managing compute, ensuring reproducibility, and integrating with CI/CD. Most teams either skip evaluation or do it poorly. Eval handles the infrastructure so you can focus on what matters: understanding your models.

10x	100%	0
Faster than DIY	Reproducible	Infrastructure to manage

Serverless runs
Submit jobs via API, get results. No infra.

Trajectory eval
Multi-turn conversations and agent paths

Model comparison
Side-by-side dashboards across models

CI/CD native
GitHub Actions, GitLab CI, webhooks

BUILT ON

rotalabs-eval open source framework. Define evaluations locally, run at scale on Eval.

INTEGRATES WITH

OpenAI, Anthropic, Cohere, Llama, Mistral, and any model with API access

01 - EVALUATION TYPES

Comprehensive LLM assessment

From single responses to complex agent trajectories, Eval provides the evaluation types you need for rigorous model assessment.

Response Evaluation

Assess individual responses against custom criteria. Define scoring functions or use pre-built evaluators.

Trajectory Evaluation

Evaluate multi-turn conversations and agent paths. Assess reasoning chains, tool usage, decision quality.

Model Comparison

Side-by-side dashboards to compare performance. Understand quality, latency, and cost tradeoffs.

Custom Evaluators

Define your own evaluation criteria and scoring functions. Bring domain-specific knowledge to your evaluations. Support for LLM-as-judge, rule-based, and hybrid approaches.

CI/CD Integration

GitHub Actions, GitLab CI, and webhook integrations. Run evaluations on every commit, PR, or deployment. Block merges that regress quality.

Parallel Execution

Serverless infrastructure scales automatically. Run thousands of evaluations in parallel without managing compute. Fast results even for large test suites.

Result Aggregation

Automatic statistical analysis and aggregation. Confidence intervals, significance testing, and trend analysis. Know when changes actually matter.

Why Reproducibility Matters

Ad-hoc evaluation scripts produce inconsistent results. When performance changes, you can't tell if it's the model or the evaluation. Eval enforces reproducibility through versioned specs, deterministic execution, and immutable result records.

Every evaluation run is reproducible. Same inputs, same outputs, every time.

02 - HOW IT WORKS

From test cases to results

Eval handles the infrastructure so you can focus on defining what matters.

01

Define

Create evaluation specs with your test cases, criteria, and scoring functions. Use YAML, JSON, or Python SDK. Version control your evaluations alongside your code.

02

Submit

Upload evaluation jobs via API or CI/CD integration. Queue jobs for immediate or scheduled execution. Tag runs for easy filtering and comparison.

03

Run

Eval executes evaluations on serverless infrastructure. Parallel execution for fast results. Automatic retries and error handling. No compute to manage.

04

Analyze

Review results in dashboards and export reports. Track trends over time. Compare across models, prompts, and configurations. Get notified on regressions.

03 - USE CASES

Where Eval delivers value

MODEL SELECTION

Choosing the Right Model

Compare GPT-4, Claude, Llama, and others on your specific use cases. Understand quality vs. cost vs. latency tradeoffs with real data from your domain.

PROMPT ENGINEERING

Prompt Optimization

Test prompt variations systematically. Know which version performs best with statistical confidence. Stop guessing, start measuring.

CI/CD

Continuous Evaluation

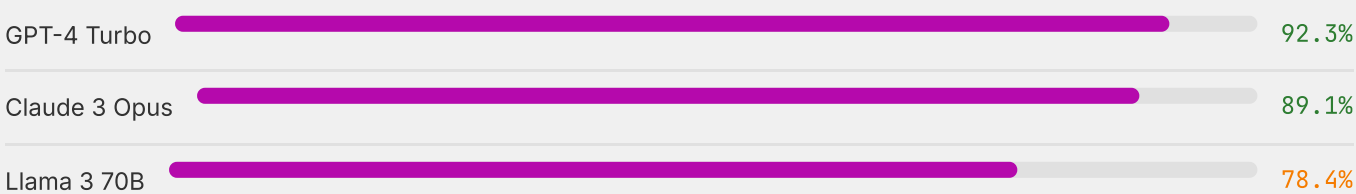
Run evaluations on every PR. Catch regressions before they hit production. Block merges that degrade quality below thresholds.

AGENT DEVELOPMENT

Agent Trajectory Testing

Evaluate multi-step agent workflows. Assess tool selection, reasoning quality, and goal completion across complex scenarios.

Example: Model Comparison Dashboard



04 - TECHNICAL SPECIFICATIONS

Enterprise-ready evaluation

CAPABILITY	SPECIFICATION
Evaluation Types	Response, trajectory, model comparison, regression testing
Supported Models	OpenAI, Anthropic, Cohere, Llama, Mistral, any API-accessible model
Parallelization	Up to 1000 concurrent evaluations per run
Custom Evaluators	Python SDK, LLM-as-judge, rule-based, hybrid
CI/CD Integration	GitHub Actions, GitLab CI, Jenkins, webhooks
Result Storage	90 days (Pro), unlimited (Enterprise), export to S3/GCS
Deployment	SaaS, private cloud (AWS, GCP, Azure), on-premise

05 - PRICING

Plans for every scale

Pay-as-you-go

$0.10/eval

Basic evaluation with no commitment. For occasional testing and experimentation.

Pro

$500/mo

10K evaluations included, scheduled runs, dashboards, CI/CD integration.

Enterprise

Custom

Unlimited evaluations, dedicated compute, on-premise deployment, custom SLA.

Pricing is indicative. Contact us for volume discounts and custom requirements.

Start evaluating today

Get your first 100 evaluations free. No credit card required.

Start Free | Learn More