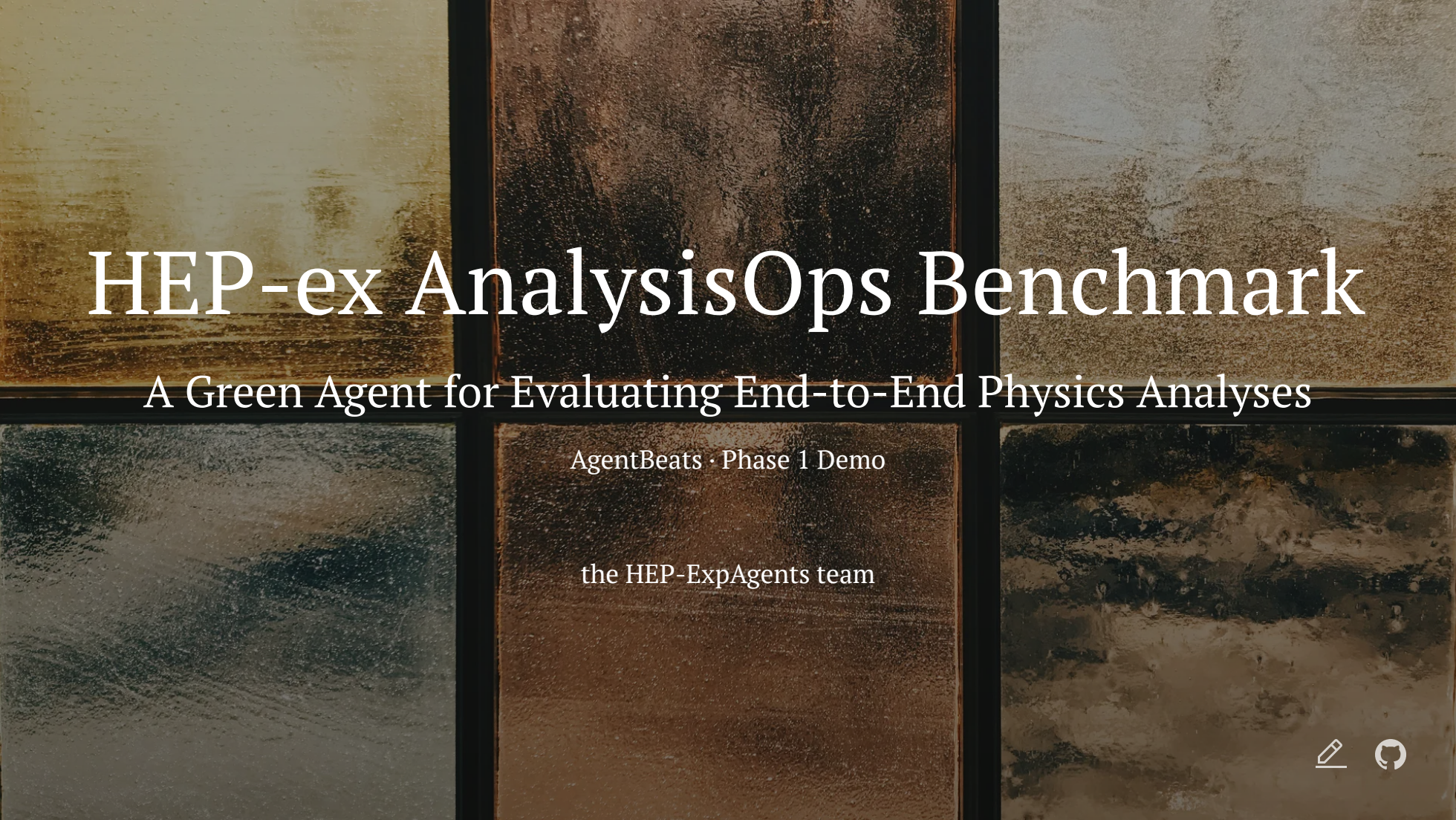
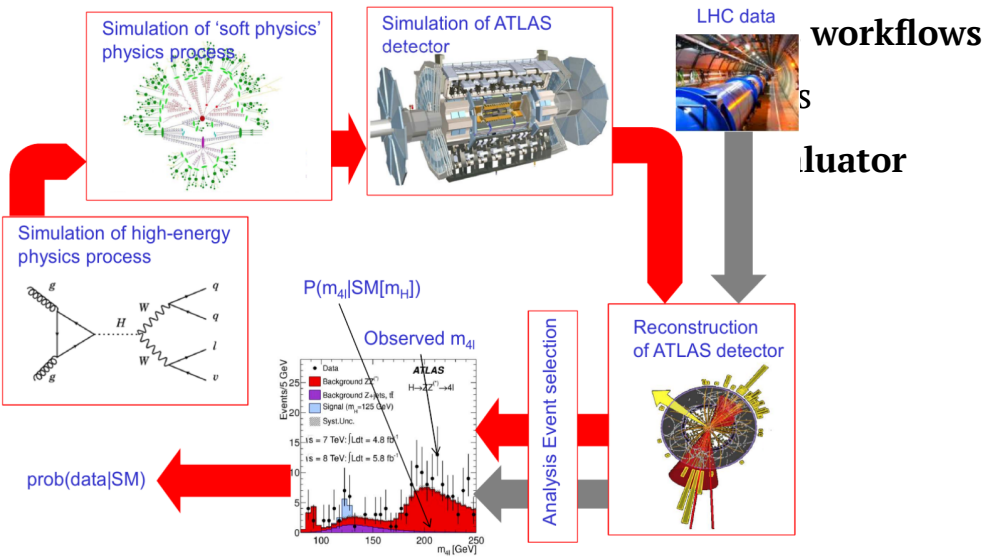


HEP-ex AnalysisOps Benchmark

A Green Agent for Evaluating End-to-End Physics Analyses

AgentBeats · Phase 1 Demo

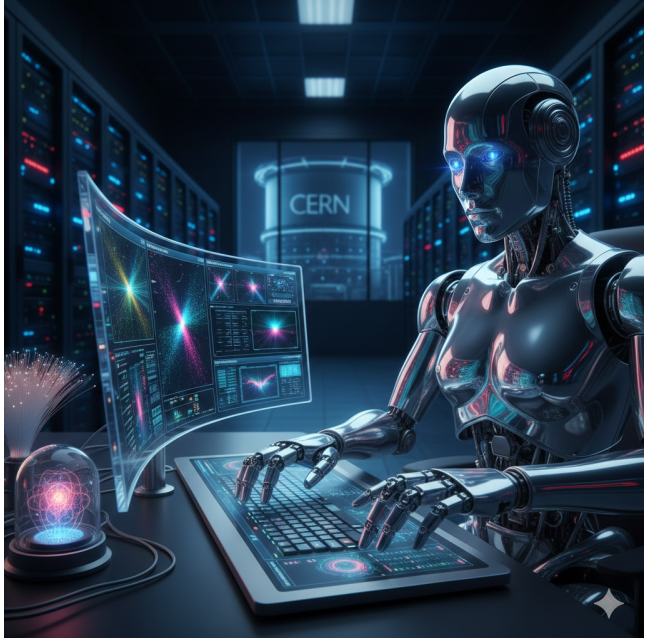
the HEP-ExpAgents team

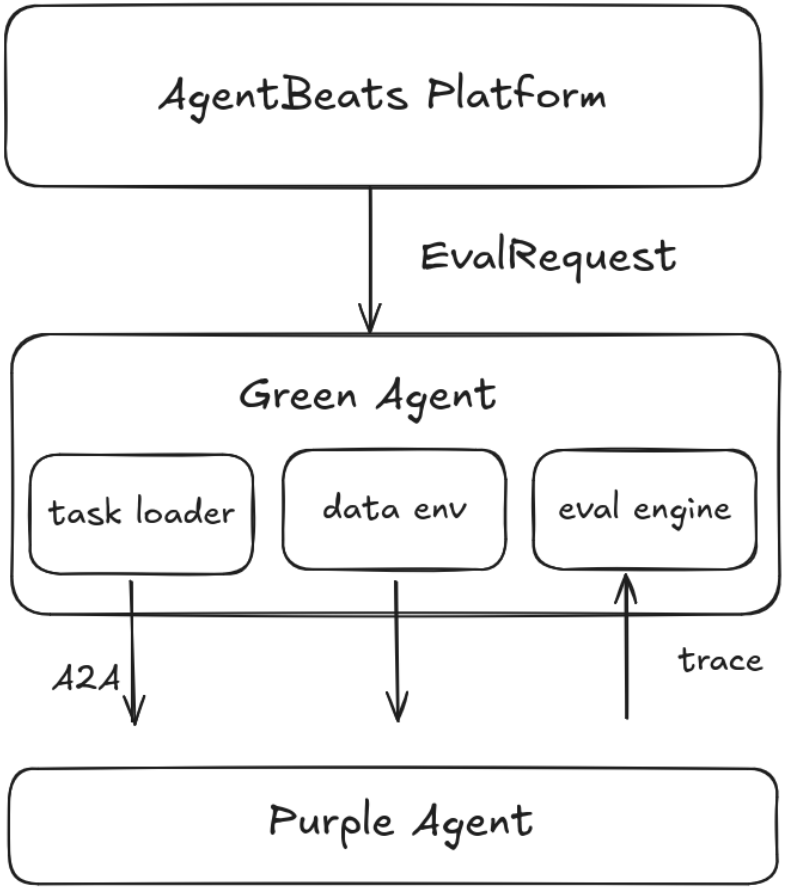


workflows

s

luator




AgentBeats Platform
EvalRequest
Green Agent
task loader
data env
eval engine
A2A
trace
Purple Agent

Evaluation Engine

```
package_loader.py
rule_engine.py
aggregator.py
```

Spec-driven evaluation

- `task_spec.yaml`

  Execution & environment contract
- `rubric.yaml`

  Three-layer scoring definition
- `eval_ref.yaml`

  Reference values / expectations
- `white_prompt.md`
- `judge_prompt.md`

Three-Layer Rubric

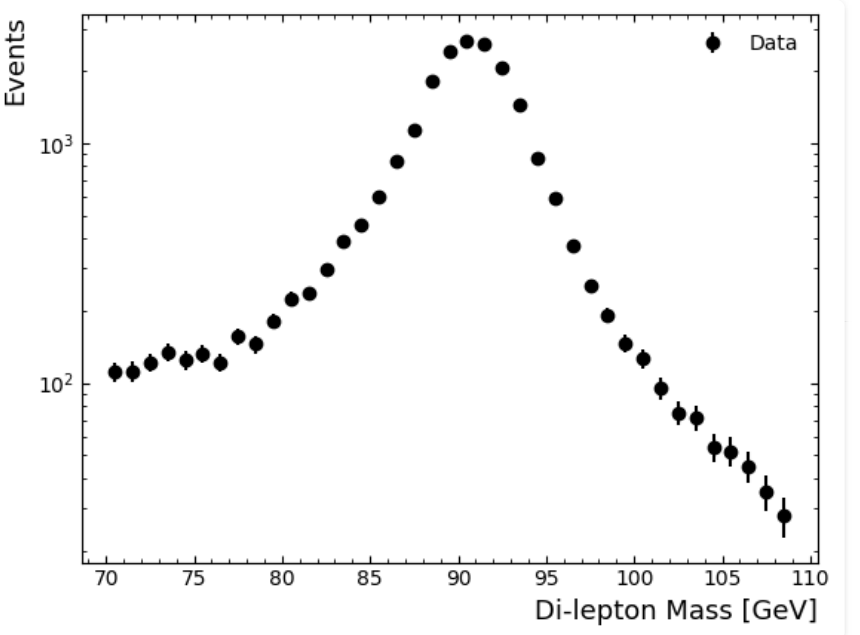
```yaml
# specs/task/rubric.yaml
# Hard correctness first
gates:
  - id: trace_present
    required_fields: ["status", "fit_result", "fit_method"]
    fail_total_score: 0

# Deterministic scoring second
rule_checks:
  - id: mu_closeness
    points: 40
    value_path: "fit_result.mu"
    ...
# Flexible scientific judgment last
llm_checks:
  - id: method_reasoning
    type: llm_reasoning
    ...
```

Core Physics Output

Target observable

- Fitted peak mean μ extracted from di-muon invariant mass
- μ used as a deterministic signal in rule-based evaluation

```
{
  "fit_result": {
    "mu": 90.75,
    "sigma": 2.28,
    ...
  }
}
```



- Scores aggregated from hard, rule-based, and LLM checks
- Structured diagnostics enable partial credit and feedback


```
{
  "status": "ok",
  "hard_checks_passed": true,

  "final": {
    "normalized_score": 1.0
  },

  "rule": {
    "score": 85.0
  },

  "llm": {
    "score": 17.0,
    "comment": "Reasonable fit strategy with minor missing details."
  },

  "issues": [
    { "severity": "warn", "message": "Missing initial fit parameters." },
    ...
  ]
}
```

- Evaluates **end-to-end physics analysis workflows**
- Combines **hard checks**, **deterministic rules**, and **LLM judgment**
- Designed as an **extensible AnalysisOps benchmark**

Looking Forward: Phase 2

- Extend to **more complex physics analyses**
- Enable **more capable agents** to demonstrate advanced workflows
- Explore **agent teams** for collaborative analysis tasks

Stay tuned.

Use code snippets and get the highlighting directly, and even types hover!

filename-example.ts

```ts
// TwoSlash enables TypeScript hover information
// and errors in markdown code blocks
// More at https://shiki.style/packages/twoslash
import { computed, ref } from 'vue'

const count = ref(0)
const doubled = computed(() => count.value * 2)

doubled.value = 2
```

```
Cannot assign to 'value' because it is a read-only
```

```ts
// Inside ./snippets/external.ts
export function emptyArray<T>(length: number) {
  return Array.from<T>({ length })
}
```

Learn more