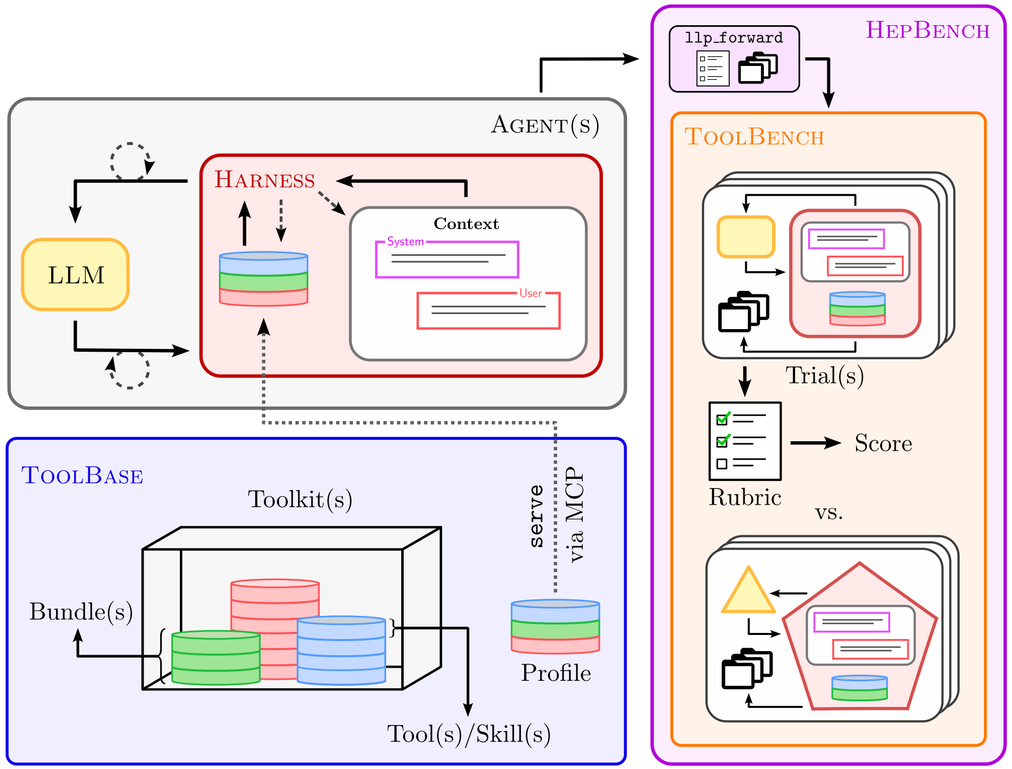
Harness HEPineering



Agentic AI in HEP Workshop @ FNAL LPC

September 11th, 2026

Based on 2609.00107, 2609.xxxx

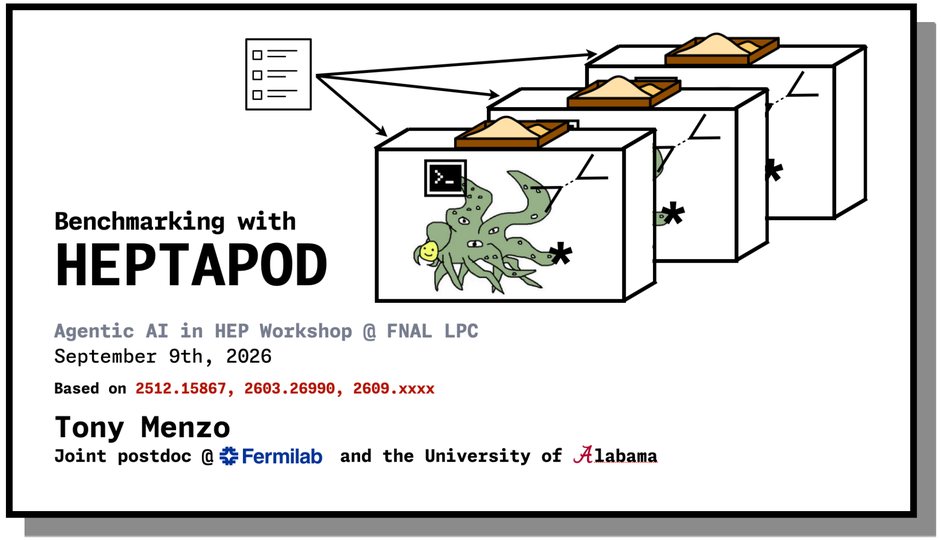
Tony Menzo

Joint postdoc @ Fermilab and the University of Alabama

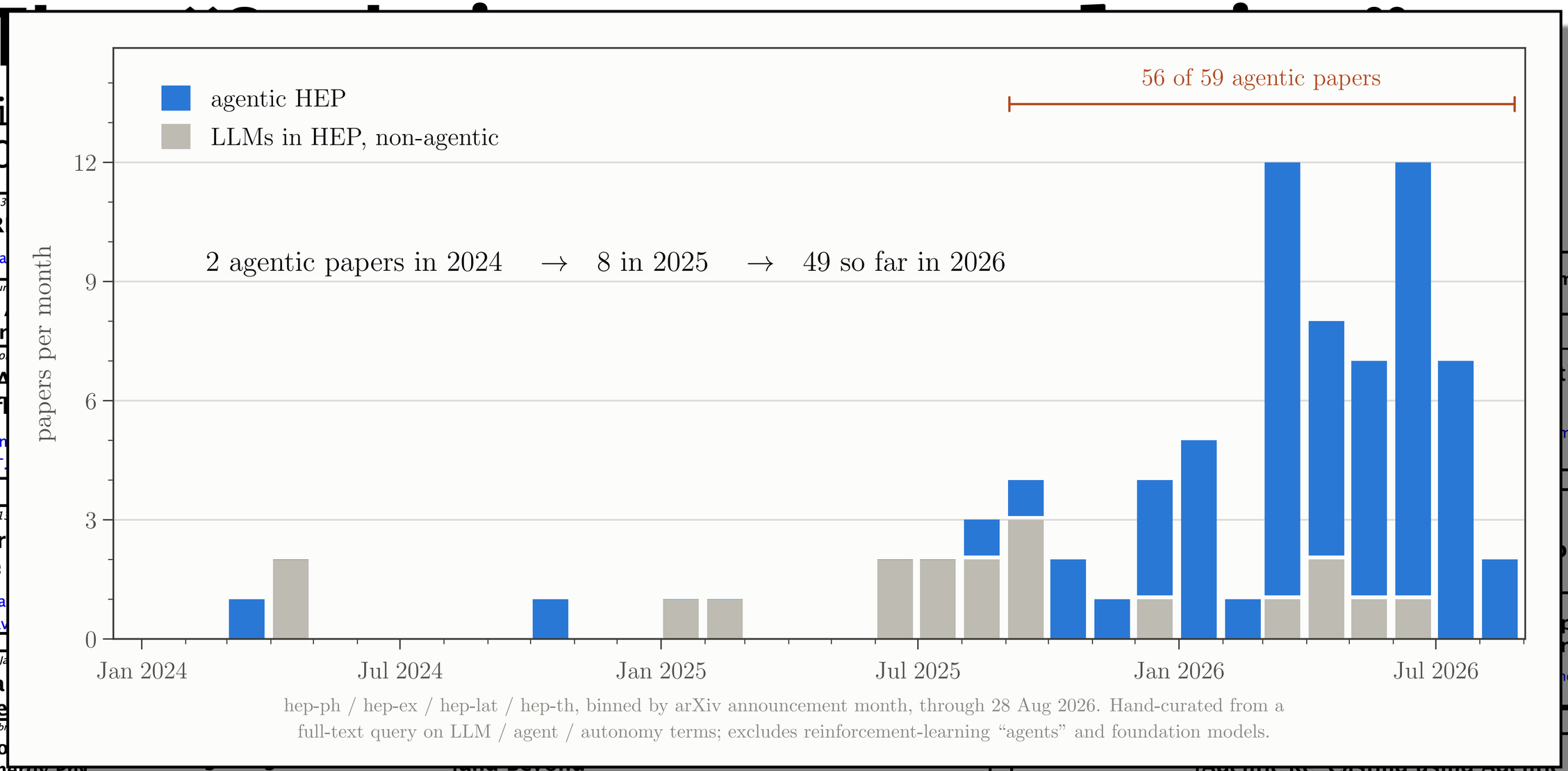
Outline

Continuing from Wednesday…

https://indico.fnal.gov/event/75639/contributions/349012/attachments/201261/280955/[Menzo]BenchmarkingWithHEPTAPOD.pdf

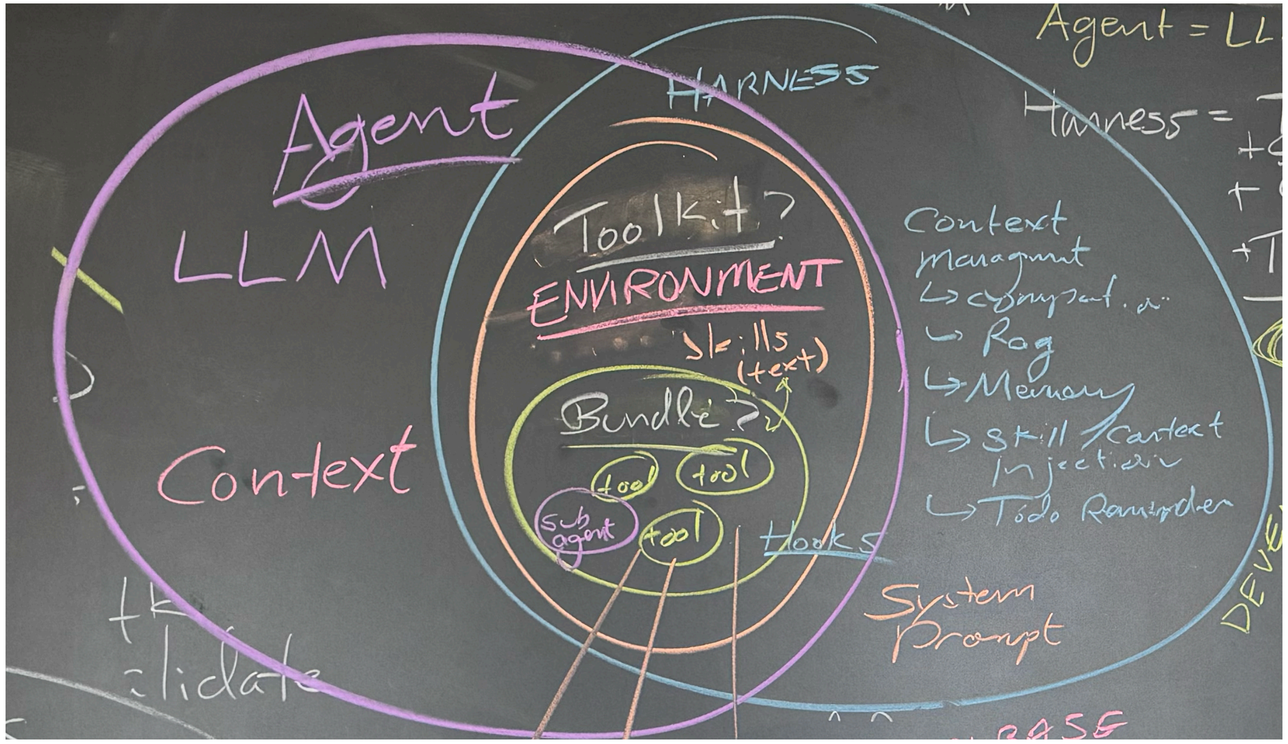


1. Status of the field (the “zoo”)
2. Toward portable/cumulative/reproducible agent science in HEP (**the *HEPTABench* stack**)
   - HEPTAPOD, toolbase, toolbench, hepbench
3. Conclusions


agentic HEP
LLMs in HEP, non-agentic
56 of 59 agentic papers
2 agentic papers in 2024 → 8 in 2025 → 49 so far in 2026
papers per month
0
3
6
9
12
Jan 2024
Jul 2024
Jan 2025
Jul 2025
Jan 2026
Jul 2026
hep-ph / hep-ex / hep-lat / hep-th, binned by arXiv announcement month, through 28 Aug 2026. Hand-curated from a full-text query on LLM / agent / autonomy terms; excludes reinforcement-learning "agents" and foundation models.

Harness?

“Everything but the LLM”



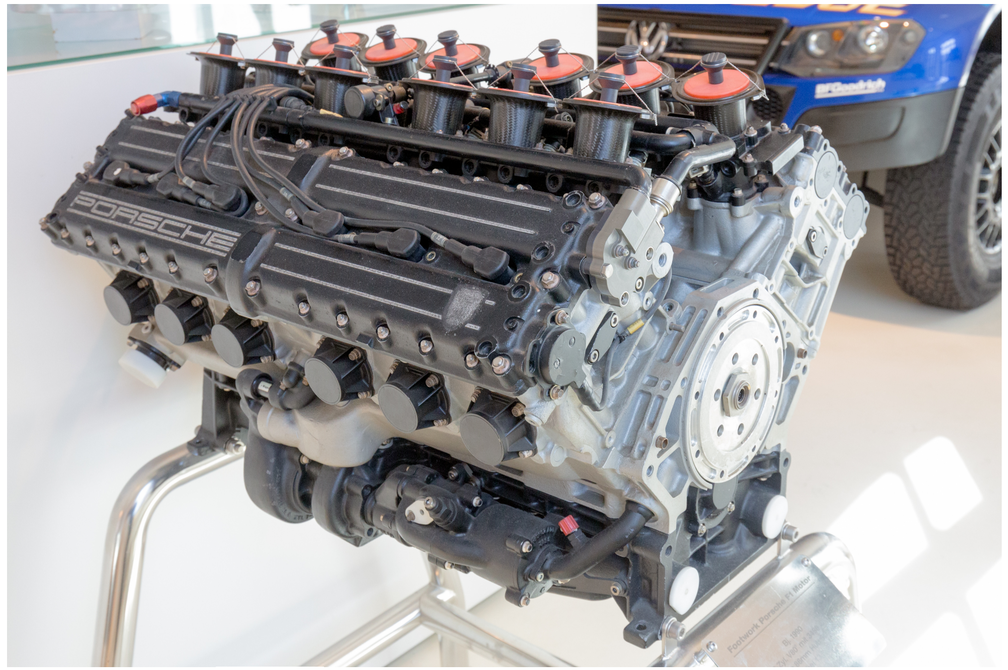
Harness?

“Everything but the LLM”

LLM

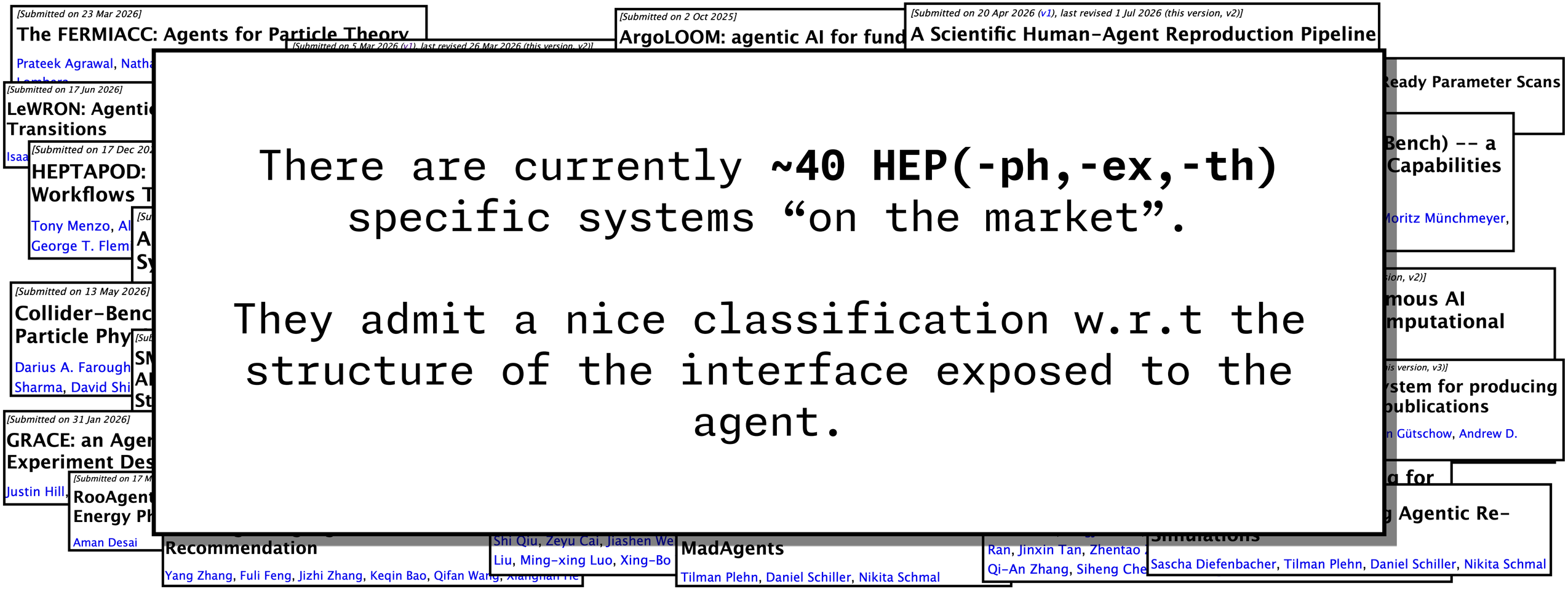


Harness



Harness design in HEP (the “zoo”)

A taxonomy



There are currently **~40 HEP(-ph,-ex,-th)** specific systems “on the market”.

They admit a nice classification w.r.t the structure of the interface exposed to the agent.

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

	Name	Architecture	MCP	Multi-agent	Domain / role
Free-form	MadAgents [4]	Free-form code generation	–	✓	Collider phenomenology
	HEP Data Analysis Agents [7]	Free-form generation in fixed workflow DAG	–	✓	Experimental analysis
	JFC [1]	Free-form generation, literature-grounded	–	✓	Experimental analysis
	CoLLM [8]	Free-form generation, structured prompt	–	✗	Collider DL analysis
	PhysMaster [9, 10]	Free-form generation, validated-trace memory	–	✓	Theory / lattice QCD
	Agents of Discovery [5]	Free-form code generation	–	✓	New-physics analysis
	ArgoLOOM [2]	Orchestrated module wrappers, RAG-grounded	–	✓	Cosmology, collider, nuclear
	ALEPH thrust [3]	Free-form generation, physicist-directed	–	✗	Experimental analysis
	SFitter Agents [6]	Orchestrator with consultant agents	–	✓	Global SMEFT re-casting
	CLVisc Agent [23]	Free-form execution, agent-authored skill	–	✗	Heavy-ion hydrodynamics
	Amplitude program search [35]	Coding agent in generate–evaluate–select loop	–	✗	Scattering amplitudes
	Detector co-design [36]	Agentic loop over bi-level optimizer	–	✗	Detector design

Free-form

Agent produces executable behavior directly through unconstrained code or command generation; domain context can guide it, but no fixed interface determines what operations it can express.

The burden of verification:

When?	Who/Where?	Per-run burden
After or during every generated implementation	Human	Highest

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

Free-form

Agent prod
behavior d
unconstrai
generation
guide it, but no fixed
interface determines what
operations it can express.

e.g. MadAgents
(https://github.com/MadGraphTeam/MadAgents/tree/main/supplementary/prompts)

	Name	Architecture	MCP	Multi-agent	Domain / role
	MadAgents [4]	Free-form code generation	–	✓	Collider phenomenology
	HEP Data Analysis Agents [7]	Free-form generation in fixed workflow DAG	–	✓	Experimental analysis
	JFC [1]	Free-form generation, literature-grounded	–	✓	Experimental analysis
	CoLLM [8]	Free-form generation, structured prompt	–	✗	Collider DL analysis
	PhysMaster [9, 10]	Free-form generation, validated-trace	–	✓	Theory / lattice QCD
					analysis
					collider, nuclear
					l analysis
					FT re-casting
					ydrodynamics
					mplitudes
	search [35]	loop			
	Detector co-design [36]	Agentic loop over bi-level optimizer	–	✗	Detector design

When?	Who/Where?	Per-run burden
After or during every generated implementation	Human	Highest

The burden of verification:

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

	Name	Architecture	MCP	Multi-agent	Domain / role
Structured handoffs	Dr. Sai [11]	Role agents over experiment software	–	✓	Experimental analysis
	DarkAgents [12]	Sub-agents with JSON handoff contracts	–	✓	Astroparticle / dark sectors
	SHARP [13]	Machine-readable project spec, subagents	–	✓	Analysis reproduction
	AgentRivet [14]	Provider-agnostic agents, output schemas	–	✓	Analysis preservation
	Semantic differencing [15]	Multi-agent trace and critique of generated code	–	✓	Analysis audit
	Selection-list codegen [16]	Structured selection list between stages	–	✓	Analysis reproduction
	LQCDMaster [17]	Hybrid: skill-guided generation, deterministic Wick tool	–	✓	Lattice QCD

Structured-handoffs

Agent generates implementations freely, but the intermediate artifacts exchanged between agents or workflow stages must follow a prescribed format or schema.

The burden of verification:

When?	Who/Where?	Per-run burden
After or during every generated implementation	Handoff validators, Human	High

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

	Name	Architecture	MCP	Multi-agent	Domain / role
Structured handoffs	Dr. Sai [11]	Role agents over experiment software	–	✓	Experimental analysis
	DarkAgents [12]	Sub-agents with JSON handoff contracts	–	✓	Astroparticle / dark sectors
	SHARP [13]	Machine-readable project spec, subagents	–	✓	Analysis reproduction
	AgentRivet [14]	Provider-agnostic agents, output schemas	–	✓	Analysis preservation
	Semantic differencing [15]	Multi-agent trace and critique of generated	–	✓	Analysis audit

Structured-handoffs

e.g. DarkAgents (https://github.com/PhysicsZandi/DarkAgents/blob/main/examples/su2conformal/claude_workspace_opus4.8/output/su2cw-dark/handoff_constraints.json)

follow a prescribed format or schema.

The burden of verification:

When?	Who/Where?	Per-run burden
After or during every generated implementation	Handoff validators, Human	High

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

Configuration/DSL mediated

	Name	Architecture	MCP	Multi-agent	Domain / role
Fixed backend	ColliderAgent [18]	CLI/template-mediated	✓	✓	Collider phenomenology
	HepScript [19]	Constrained-grammar DSL	–	✗	Experimental analysis
	EasyScan_HEP 2 [20]	Agent-written configuration, fixed backend	–	✗	Parameter scans
	SMEFT-Pheno-Agent [21]	Locked configuration, phase manifests	–	✗	SMEFT phenomenology

Agent specifies a workflow using a bounded configuration language, template, command interface, or domain-specific-language, which a fixed backend interprets and executes.

The burden of verification:

When?	Who/Where?	Per-run burden
Mainly when backend/ DSL is authored, then when each spec is produced	Fixed backend, parser	Moderate

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

	Name	Architecture	MCP	Multi-agent	Domain / role
Fixed backend	ColliderAgent [18]	CLI/template-mediated	✓	✓	Collider phenomenology
	HepScript [19]	Constrained-grammar DSL	–	✗	Experimental analysis
	EasyScan_HEP 2 [20]	Agent-written configuration, fixed backend	–	✗	Parameter scans
	SMEFT-Pheno-Agent [21]	Locked configuration, phase manifests	–	✗	SMEFT phenomenology

Configuration/DSL mediated

Agent specifies a workflow

e.g. HEPScript (https://arxiv.org/pdf/2605.01423 supplement)

language, which a fixed backend interprets and executes.

The burden of verification:

When?	Who/Where?	Per-run burden
Mainly when backend/ DSL is authored, then when each spec is produced	Fixed backend, parser	Moderate

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

Tool-constrained

The agent selects and composes discrete, fixed, versioned scientific operations through structured invocation structure.

	Name	Architecture	MCP	Multi-agent	Domain / role
Valid-by-rejection	RooAgent [25]	Typed tool registry over LangGraph and MCP	✓	✗	Experimental analysis
	FERMIACC [24]	Typed schemas over a guarded pipeline	–	✓	Particle theory
	GRACE [26]	Tool-invocation DAG; value-level verification against physics constraints	–	✗	Detector / experiment design
	bsm-agent [27]	Tool calling over a symbolic backend	–	✗	BSM model building
	LeWRON [33]	Audited toolbox with auditor agents	–	✓	Particle cosmology
	Archi [34]	ReAct agent over corpus and live MCP tools	✓	✗	Experimental operations
Valid-by-construction	HEPTAPOD (event generation) [31]	Typed schema constrains decoding	✓	✗	Collider phenomenology
	HEPTAPOD (Diagrammatica) [32]	Typed schema constrains decoding	✓	✗	Symbolic computation

The burden of verification:

When?	Who/Where?	Per-run burden
Tool authoring and each invocation/ composition	Tool selection/ arguments, tool call validator	Low

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

Tool-co

**The agent s
discrete, t
scientific
structured
structure.**

```python
@define_tool(param_descriptions={
    "px": "Momentum along the x axis, in GeV",
    "py": "Momentum along the y axis, in GeV",
})
def transverse_momentum(px: float, py: float) -> str:
    """
    Compute the transverse momentum pT = sqrt(px^2 + py^2) of a particle.

    Use this when you have the Cartesian momentum components of a particle and
    need its momentum in the plane transverse to the beam.

    Args:
        px: Momentum along the x axis, in GeV.
        py: Momentum along the y axis, in GeV.

    Returns:
        A one-line string with the transverse momentum in GeV.
    """
    pt = math.sqrt(px**2 + py**2)
    return f"pT = {pt:.3f} GeV"
```

omain / role
Experimental analysis
Particle theory
Detector / experiment design
BSM model building
Particle cosmology
Experimental operations
Collider phenomenology
Symbolic computation

The burd

un burden

composition	call validator	Low

Harness design in HEP (the “zoo”)

A taxonomy

https://arxiv.org/abs/2609.00107v1

```python
@define_tool(param_descriptions={
    "px": "Momentum along the x axis, in GeV",
    "py": "Momentum along the y axis, in GeV",
})
def transverse_momentum(px: float, py: float) -> str:
    """
    py: Momentum along the y axis, in GeV.

    Returns:
        A one-line string with the transverse momentum in GeV.
    """
    pt = math.sqrt(px**2 + py**2)
    return f"pT = {pt:.3f} GeV"
```

Build your first tool (how to):
https://github.com/tonymenzo/heptapod/tree/main/examples/primer

Tool-co

The
disc
scie
stru
stru

The burd

omain / role
Experimental analysis
Particle theory
Detector / experiment design
BSM model building
ogy
perations
menology
utation

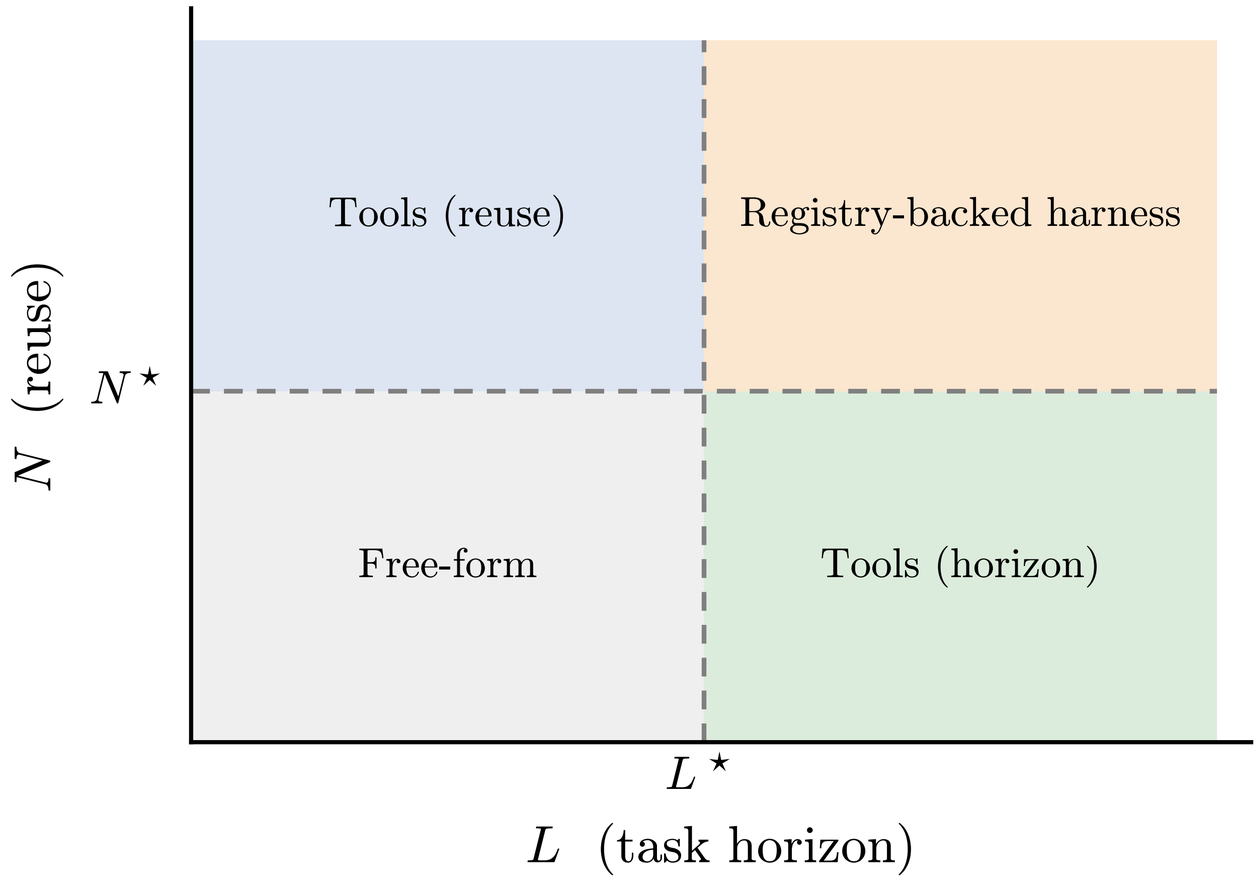
un burden

Low

composition | call validator

Harness design in HEP (the “zoo”)

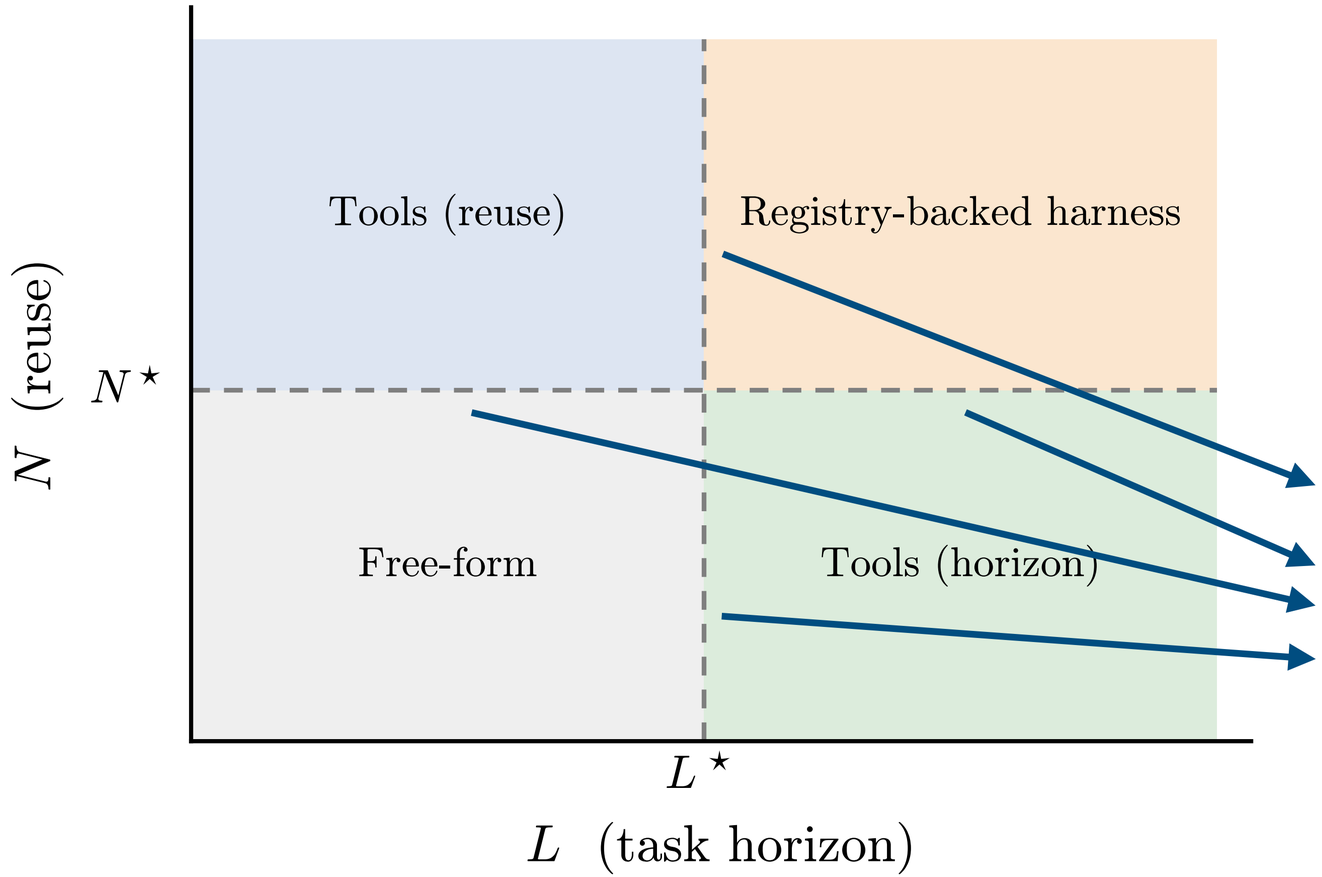
A phase diagram



*Other relevant axes, $/token, $/performance, tokens/performance, generation time, etc.

Harness design in HEP (the “zoo”)

A phase diagram



*Other relevant axes, $/token, $/performance, tokens/performance, generation time, etc.

Where are these boundaries? How can they be determined systematically?

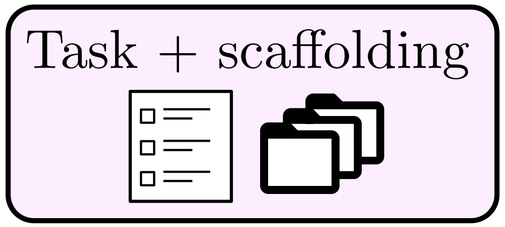
Testing the harness

“Is this thing on?”

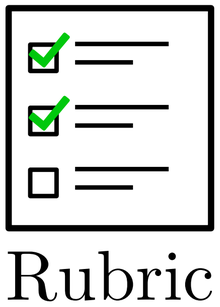
Benchmarks!

Naive minimal spec:

- **Task + task scaffolding**
  - Sandbox and workspace initialization, system prompt, etc.



- **Rubric**
  - Final expected outcome. Grading: binary or continuous, stage-by-stage grading

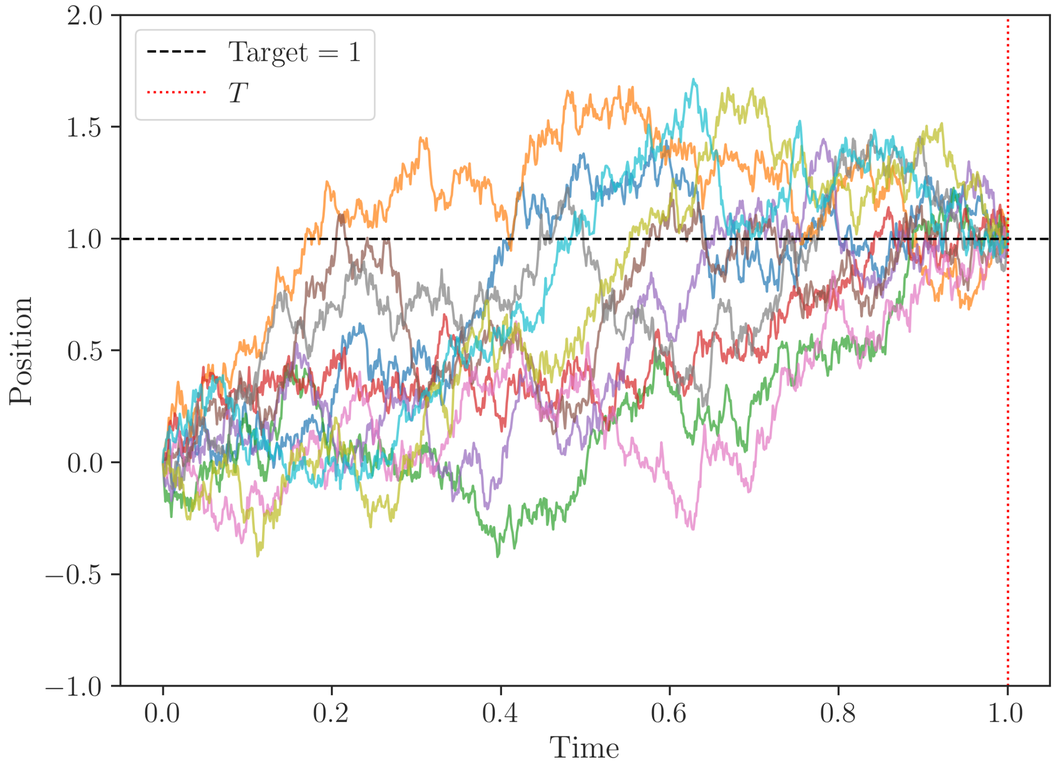


Testing the harness

"Is this thing on?"

LLMs are distributions!

$$\text{LLM} \sim p_\theta(x_{t+1} \mid x_{\leq t})$$



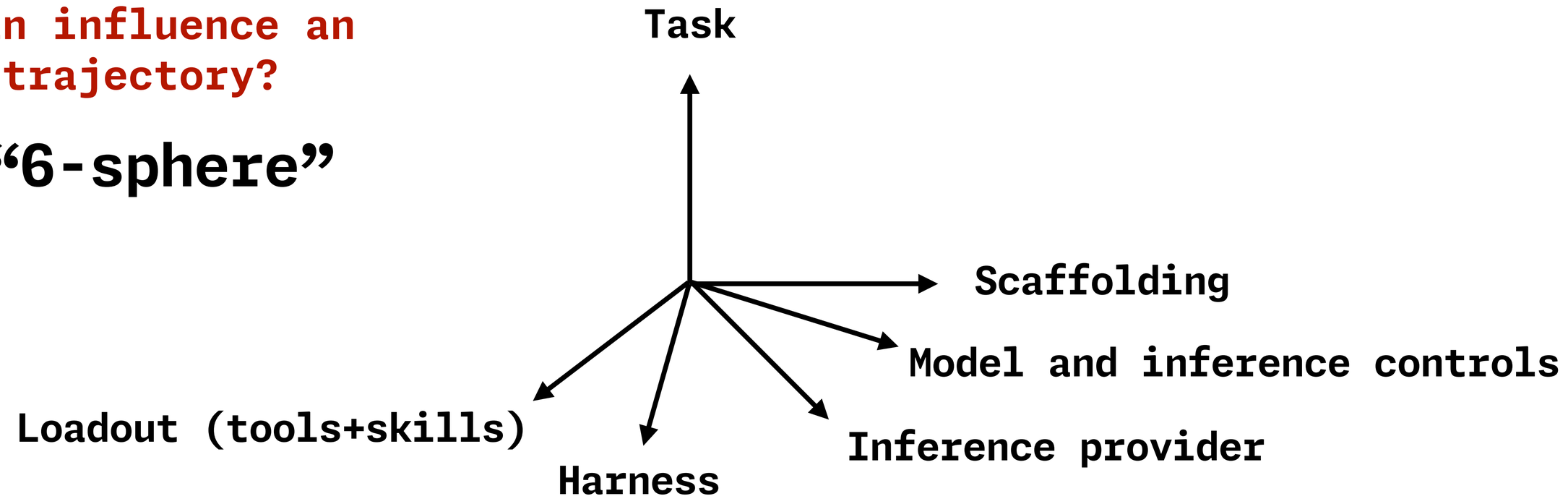
Agents trajectory/final result should be treated as a random variable with mean and variance (when applicable).

Testing the harness

"Is this thing on?"

What can influence an agents trajectory?

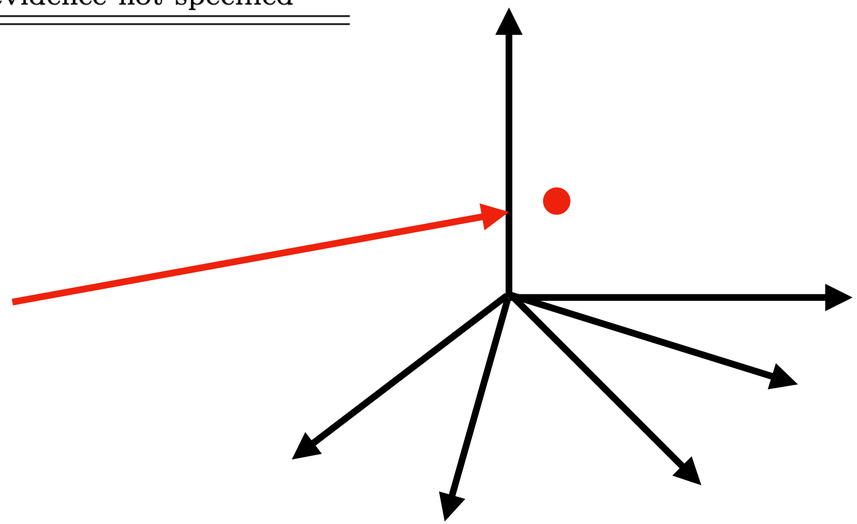
The "6-sphere"



Benchmarking in HEP

Name	Scope	Grading	Integrity control / retained evidence
Collider-Bench [22]	LHC analysis reproduction; 10 tasks from 4 papers	Histogram fidelity, no hand-written rubric; 3 repeats; cost reported	LLM judge audits workspace and session trace for fabrication
TPBench [37]	HEP theory and cosmology; 57 problems	Automated numeric verification; holistic grading of reasoning	Public/private split maintained against leakage
CritPt [38]	Frontier physics research incl. HEP; 71 challenges, 190 checkpoints	Physics-format-aware auto-grading; reliability metric; cost reported	Unpublished problems, guess-resistant by construction
PRL-Bench [39]	End-to-end physics research; 100 papers across 5 subfields	Objective verifiability; expert-validated tasks	Published-source tasks; retained execution evidence not specified

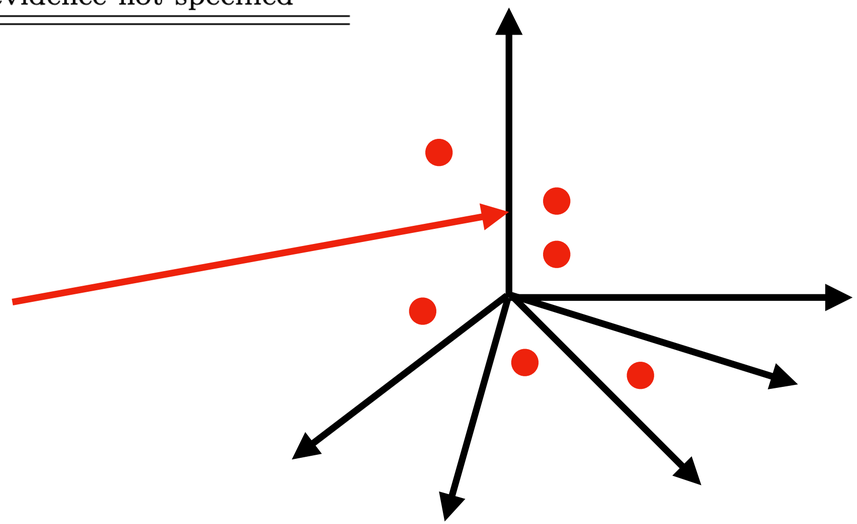
Each of these benchmarks only operate at a single point in the 6-sphere



Benchmarking in HEP

Name	Scope	Grading	Integrity control / retained evidence
Collider-Bench [22]	LHC analysis reproduction; 10 tasks from 4 papers	Histogram fidelity, no hand-written rubric; 3 repeats; cost reported	LLM judge audits workspace and session trace for fabrication
TPBench [37]	HEP theory and cosmology; 57 problems	Automated numeric verification; holistic grading of reasoning	Public/private split maintained against leakage
CritPt [38]	Frontier physics research incl. HEP; 71 challenges, 190 checkpoints	Physics-format-aware auto-grading; reliability metric; cost reported	Unpublished problems, guess-resistant by construction
PRL-Bench [39]	End-to-end physics research; 100 papers across 5 subfields	Objective verifiability; expert-validated tasks	Published-source tasks; retained execution evidence not specified

It would be ideal to scan larges parts of this space easily



Status of agents in HEP

Bottom line – it's the Wild West, as expected

As with all new fields of research, community standards are lagging behind.

Initial systems/demonstrations are promising, but hard to reuse and reproduce results without significant effort.

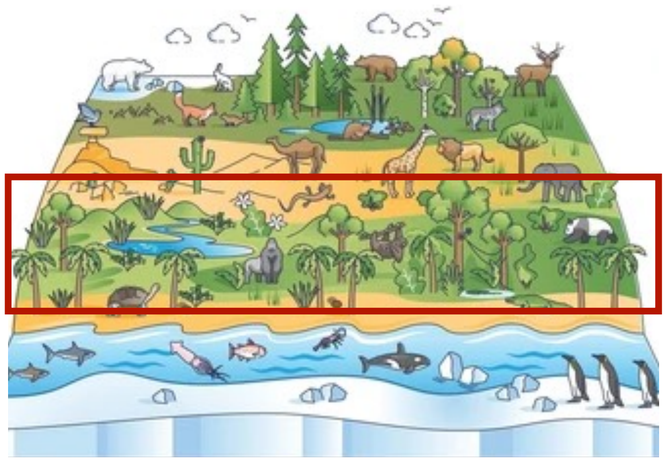
Initial benchmarks are also great, but again, reproducibility is an issue.

A cohesive framework is lacking.

Infrastructure mode… engage



Zoo

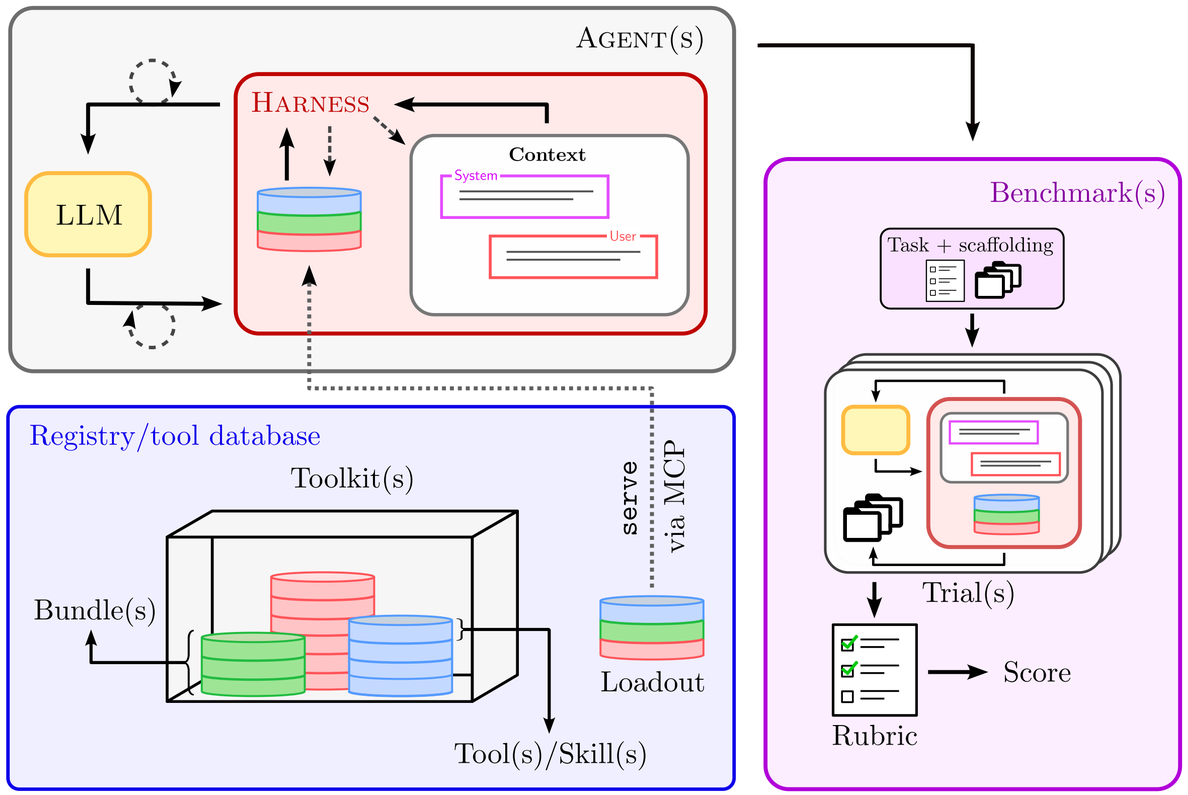


Ecosystem

Infrastructure for a community-driven HEP harness

Utilize and adapt existing infrastructure as much as possible!
General purpose coding harnesses + domain tools & skills = HEP agent

- **Adopt a standard protocol for interfacing tools with harnesses**
  - **Model Context Protocol (MCP)** has been widely adopted by all harnesses, let's not reinvent the wheel!
- **Community maintained registry/tool-skill database + serving**
  - Scalable — what do you do when you have $\mathcal{O}(1000)$ tools to choose from? Need easy curation of custom "loadouts"
- **Science-forward benchmarking framework**
  - Emphasis on reproducibility and full "6-sphere" resolution

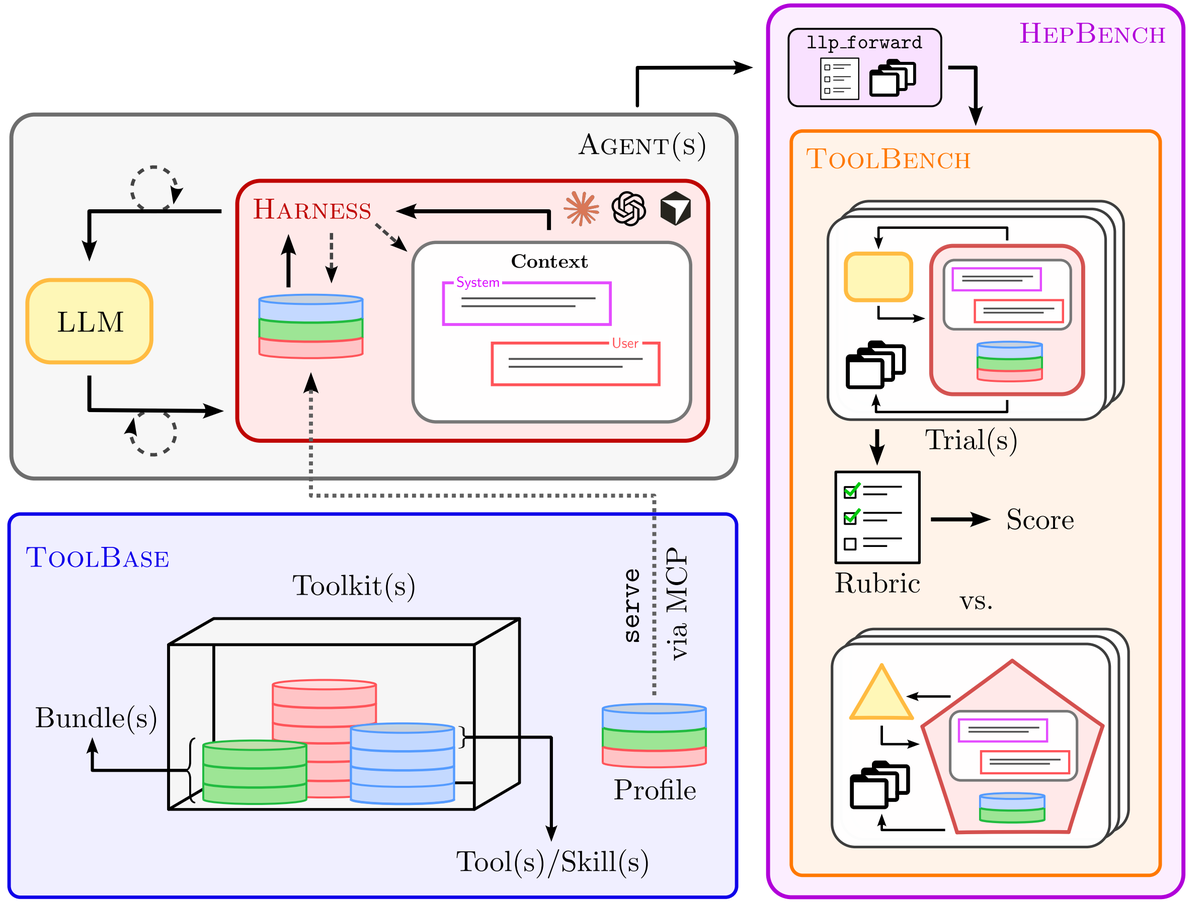


Infrastructure for a community-driven HEP harness

An implementation teaser: “the stack”

See benchmarks in Wednesday’s presentation…

- Adopt a standard protocol for interfacing tools with harnesses
  - MCP
- Community maintained registry/ tool-skill database + serving
  - toolbase
- Science-forward benchmarking framework
  - toolbench, hepbench
- Toolkit
  - heptapod

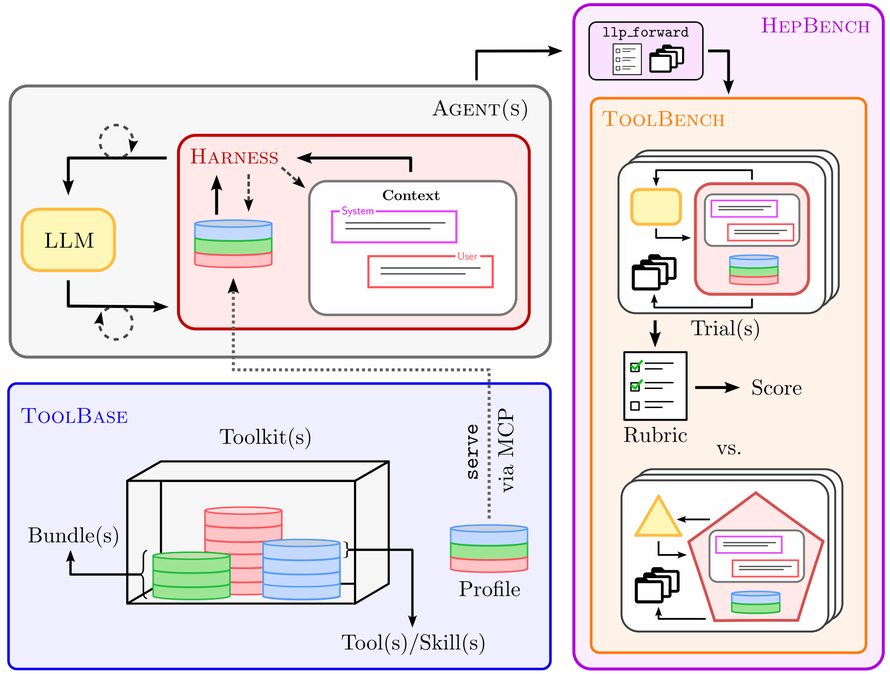


Conclusions

Harness HEPineering

https://github.com/tonymenzo/heptapod

- Agentic systems are beginning to address all corners of HEP
- With lacking community standards, results are hard to reuse, repurpose, reproduce
- The “stack” (coming soon…) (MCP + toolbase + toolbench + hepbench + heptapod) is a concrete scalable framework for testing/ maintaining future agentic systems

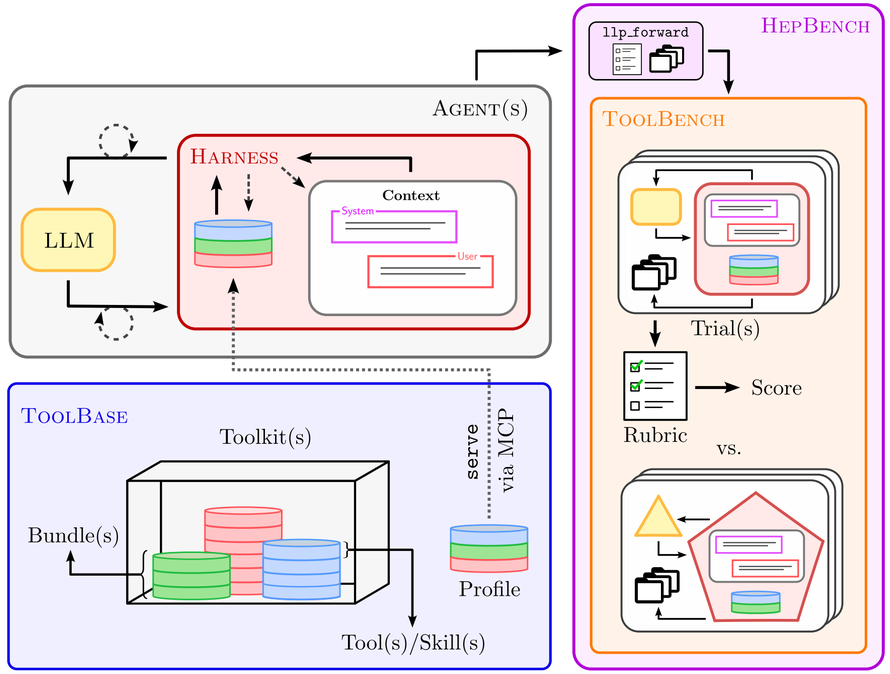


Conclusions

Harness HEPineering

https://github.com/tonymenzo/heptapod

- Agentic systems are beginning to address all corners of HEP
- With lacking community standards, results are hard to reuse, repurpose, reproduce
- The “stack” (coming soon…) (MCP + toolbase + toolbench + hepbench + heptapod) is a concrete scalable framework for testing/ maintaining future agentic systems



Thanks for your attention :)

Back-ups

Deviant behavior

