



ML4Jets 2026 · Vienna
Machine Learning for Jets and HEP

Context-Engineered Agents for High Energy Physics

D. Faroughy · S. Palacios Schweitzer · I. Pang
D. Shih · AU
Rutgers University



RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

What actually makes an agent?

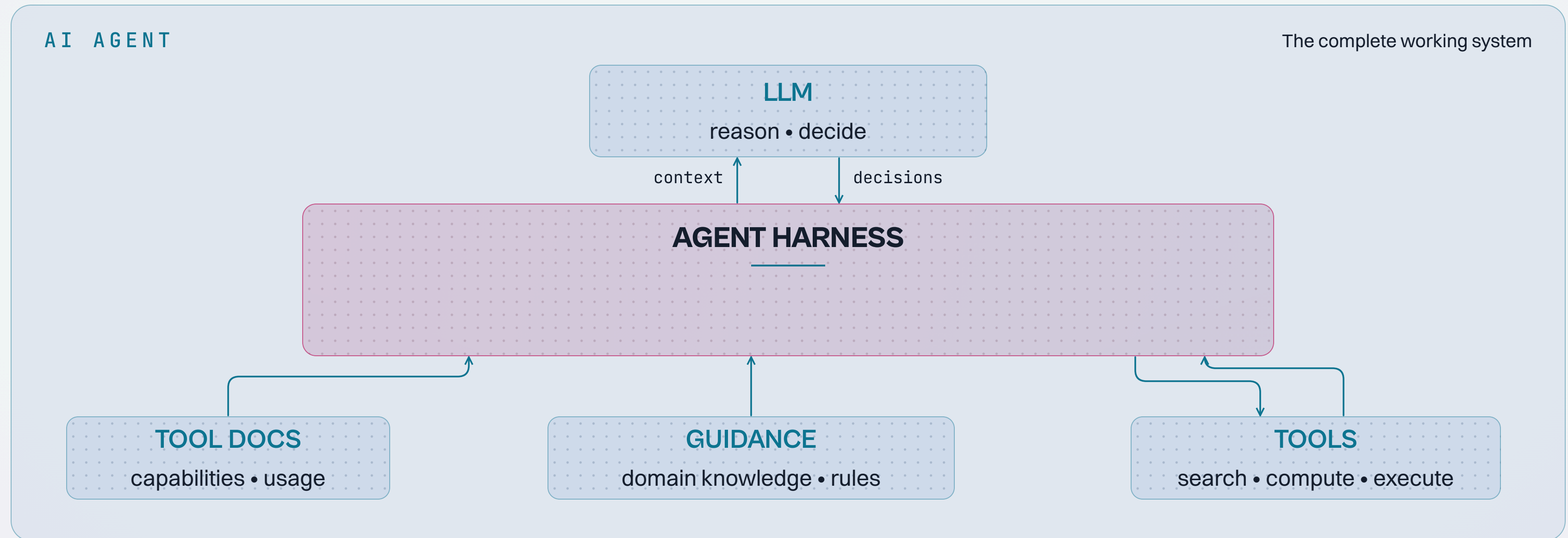
AI AGENT

The complete working system

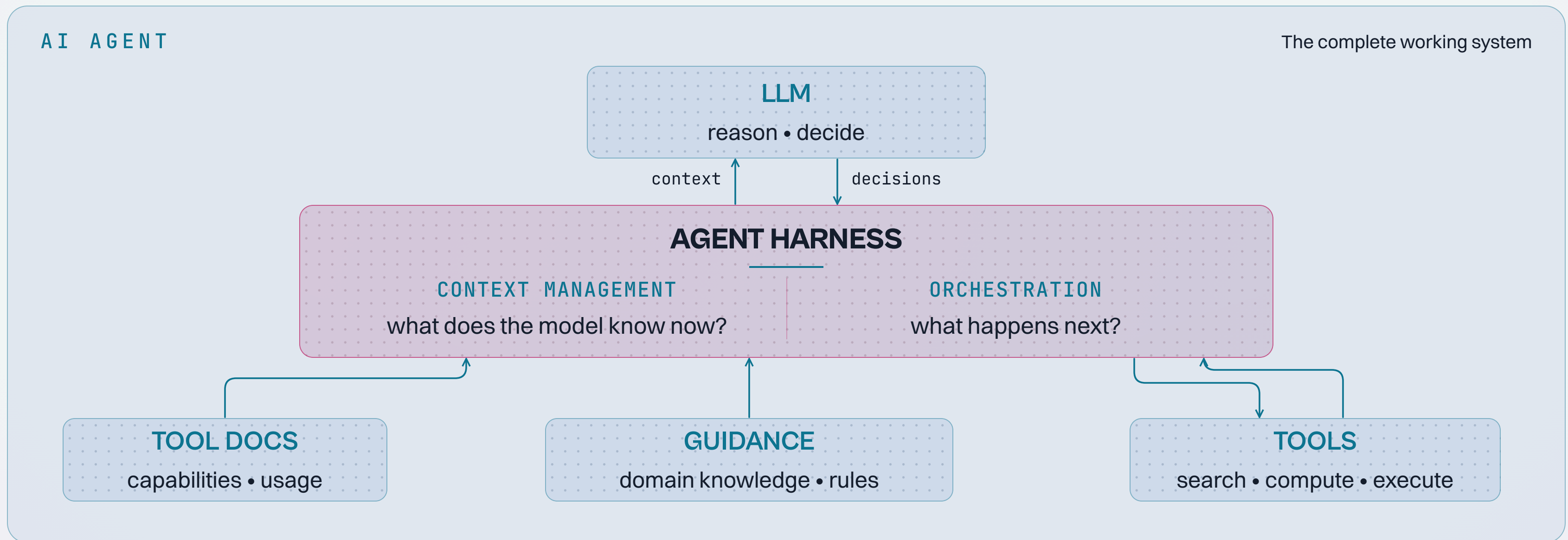
LLM

reason • decide

What actually makes an agent?

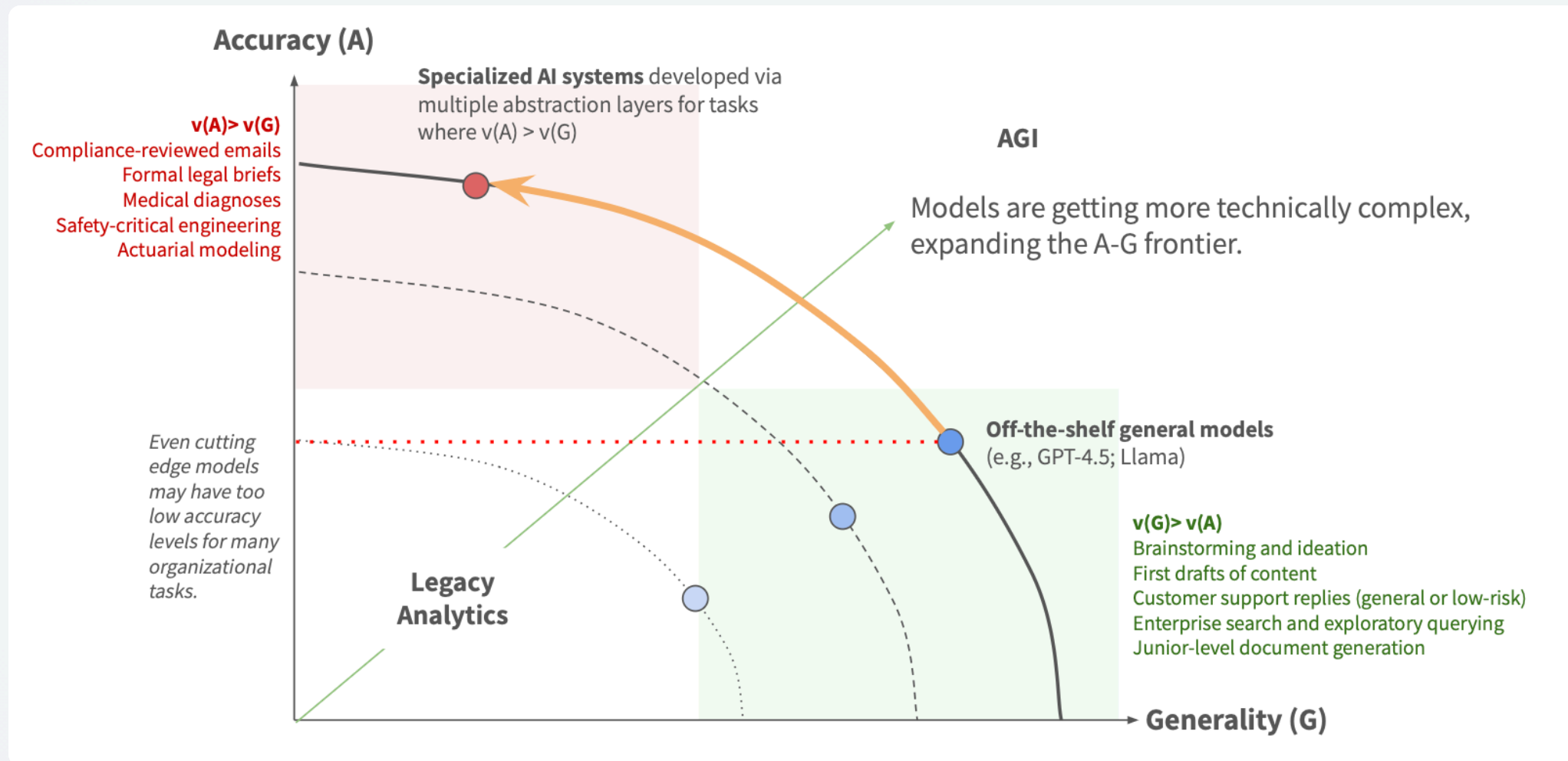


What actually makes an agent?



A strong LLM and a good harness are **both critical components** of an AI agent.

The Generality vs. Specificity Tradeoff



Hasan, Oettl & Samila · [arXiv:2506.22440](https://arxiv.org/abs/2506.22440) ↗

- There is tension between **generality and specialization**: where you choose to sit on this curve depends on the agent's requirements.
- **Borman et al. [arXiv:2607.14456]** compared specialized agents with general-purpose agents and reported that the specialized agents achieved:
 - **9–20 percentage-point improvement** in tool-use exactness.
 - **3× fewer** tool-call errors.
 - **Over 95% lower** generation token cost.

If the goal is **higher accuracy with lower token use**, and a single-field focus is acceptable, specialization may be the answer.

Ariadne

What we aim to improve

Consistently, across the full workflow.

 Long-horizon coherence

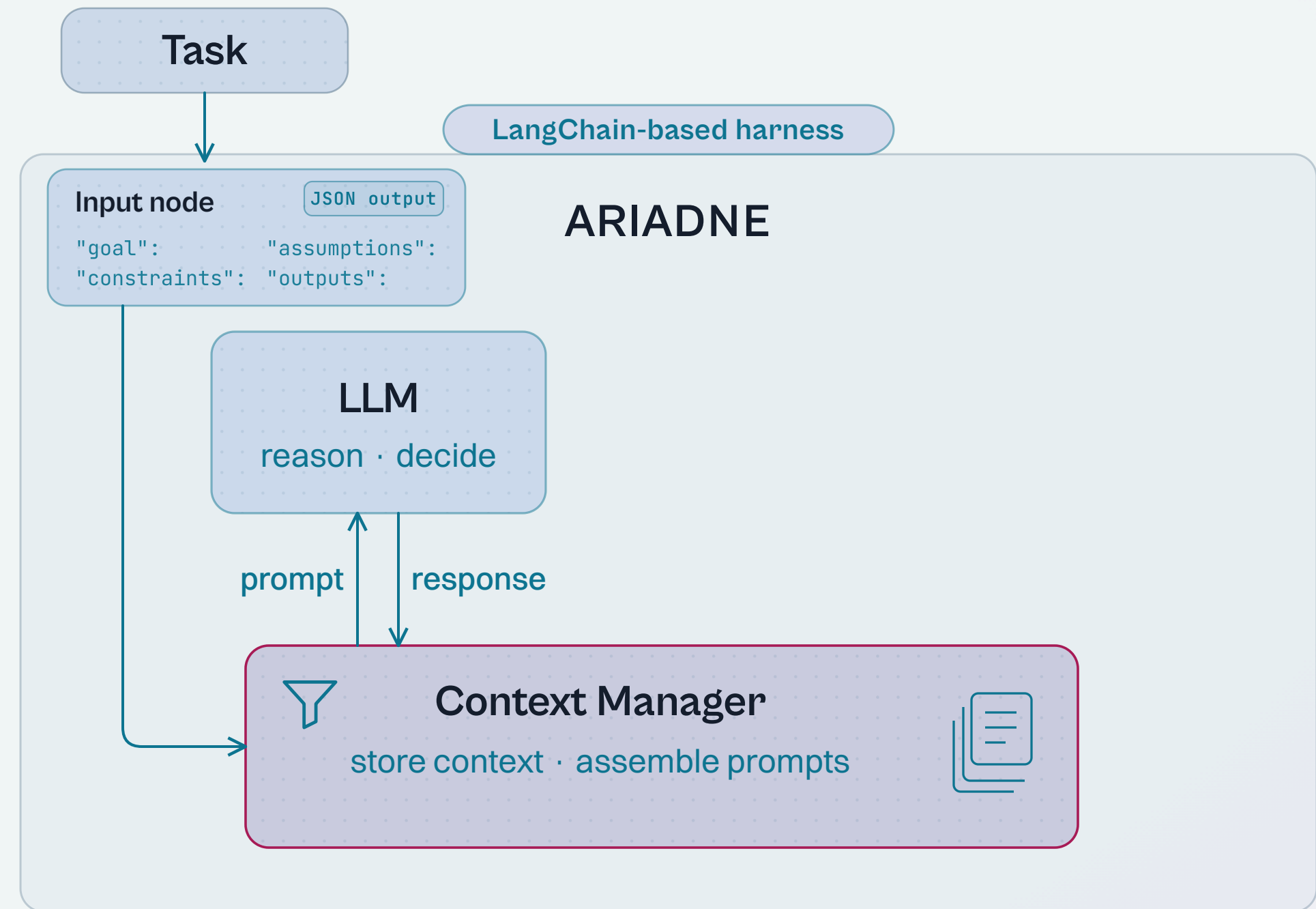
 Cost-efficient execution

 Reproducible and reliable

 LLM-agnostic

[1] Significant Gravitas (2023) · *Auto-GPT* · [GitHub](#) ↗

[2] OpenAI (2023) · *Function Calling and Other API Updates* · [13 June](#) ↗





What we aim to improve

Consistently, across the full workflow.

 Long-horizon coherence

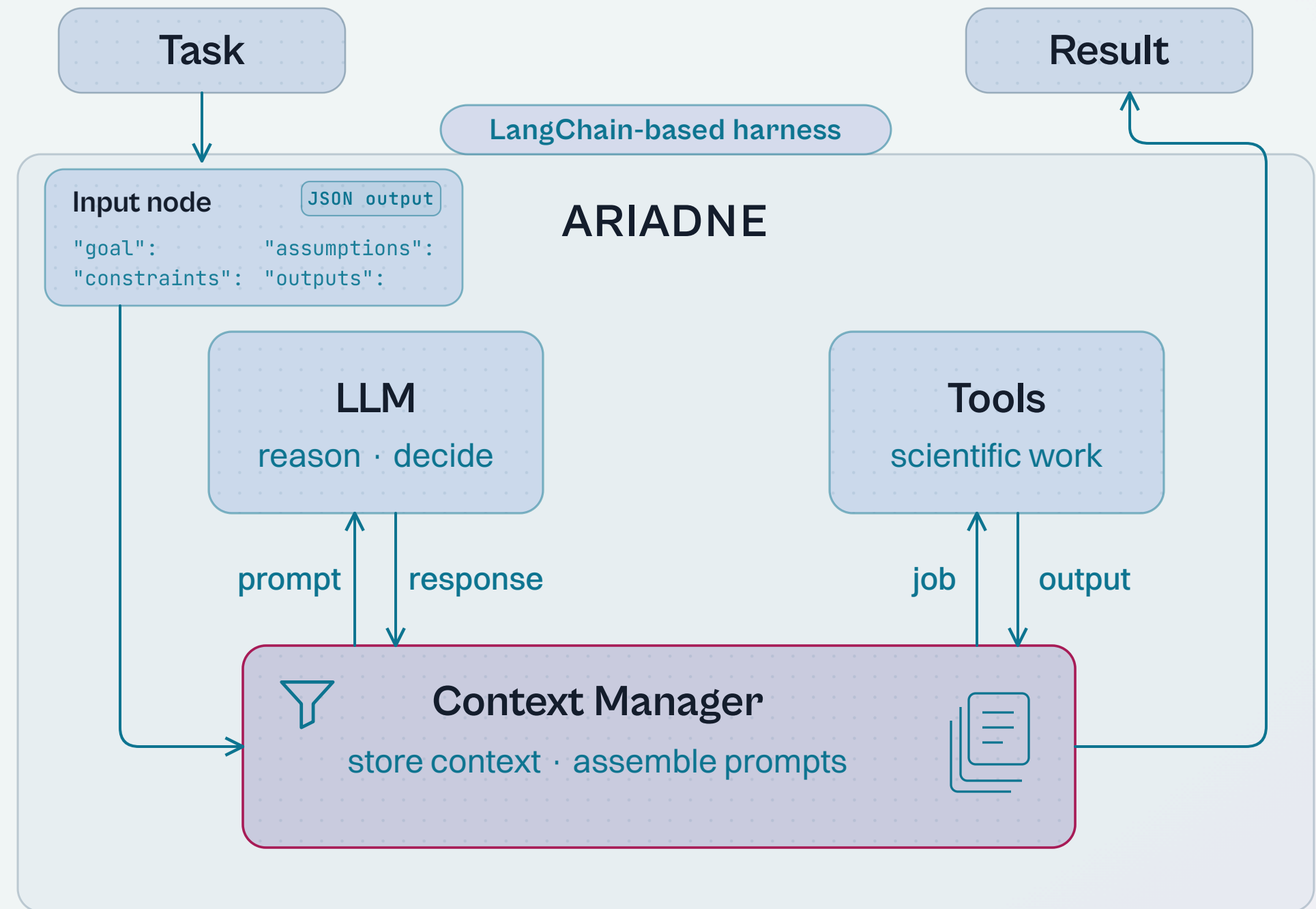
 Cost-efficient execution

 Reproducible and reliable

 LLM-agnostic

[1] Significant Gravitas (2023) · *Auto-GPT* · [GitHub](#) ↗

[2] OpenAI (2023) · *Function Calling and Other API Updates* · [13 June](#) ↗



Custom Context Management for High Energy Physics

A task that runs a full simulation pipeline · one step of it, and what crosses into the next

... earlier steps in the plan ↓

THE GLOBAL CONTEXT

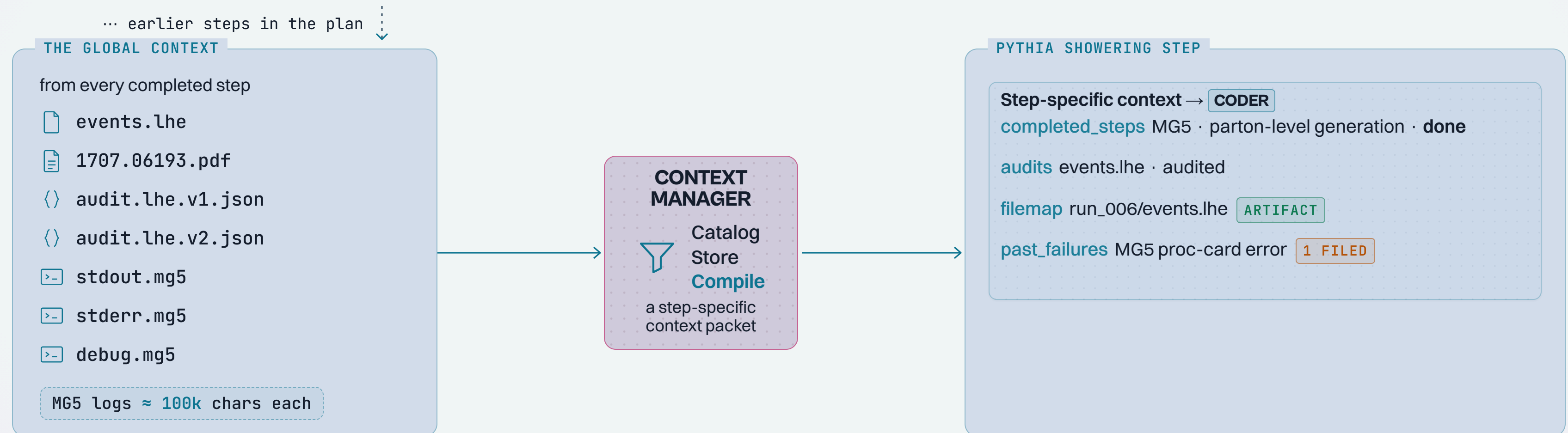
from every completed step

- events.lhe
- 1707.06193.pdf
- audit.lhe.v1.json
- audit.lhe.v2.json
- stdout.mg5
- stderr.mg5
- debug.mg5

MG5 logs ≈ 100k chars each

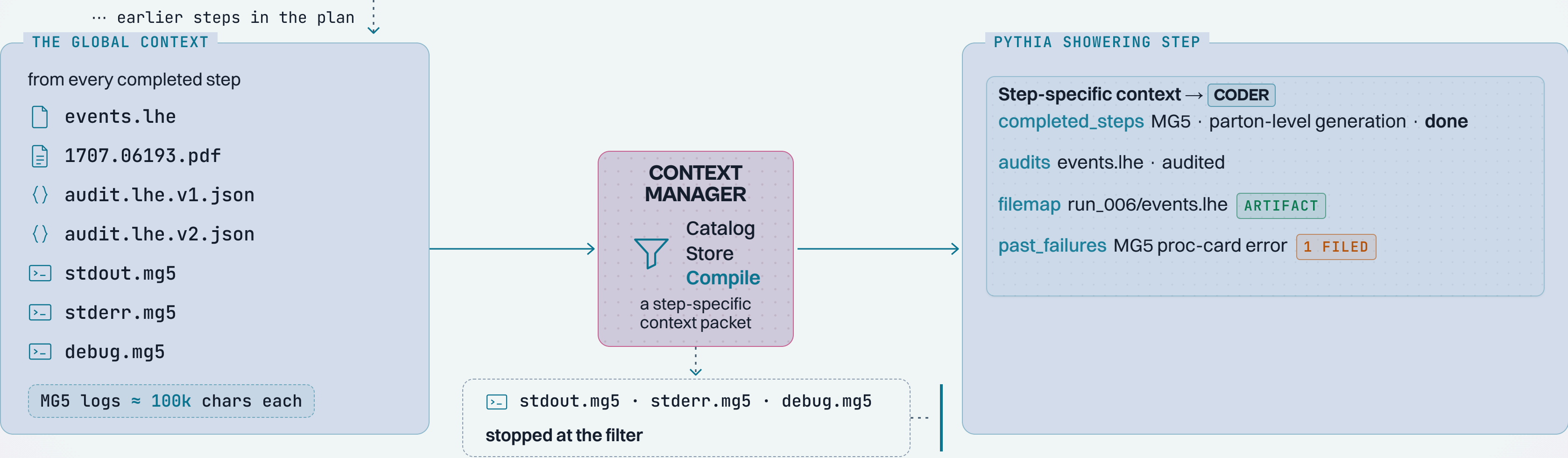
Custom Context Management for High Energy Physics

A task that runs a full simulation pipeline · one step of it, and what crosses into the next



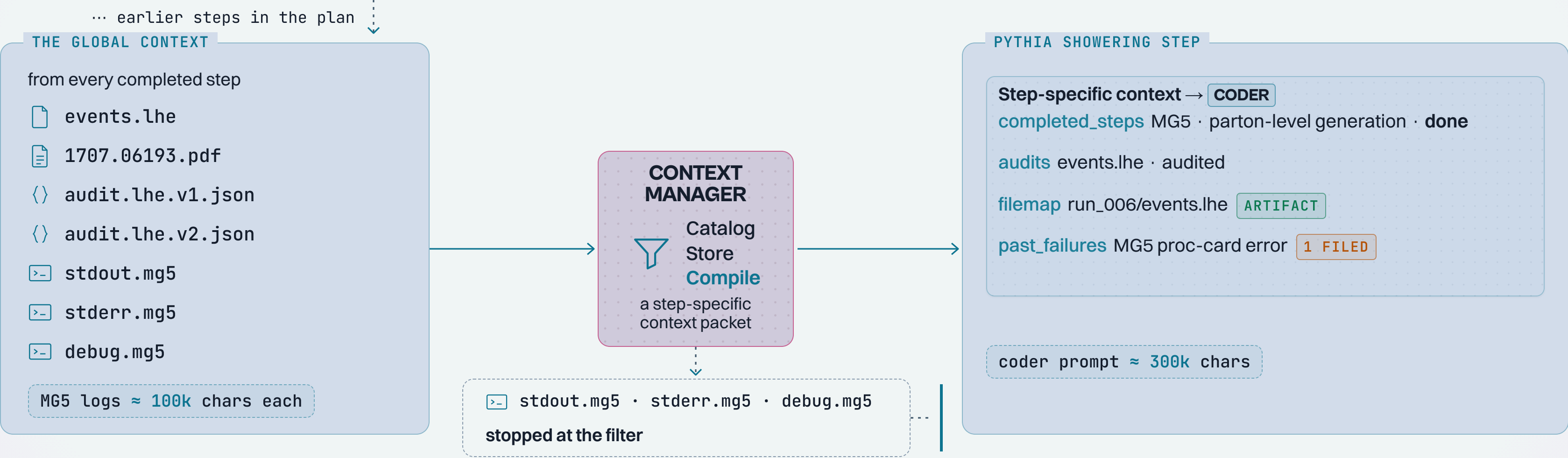
Custom Context Management for High Energy Physics

A task that runs a full simulation pipeline · one step of it, and what crosses into the next



Custom Context Management for High Energy Physics

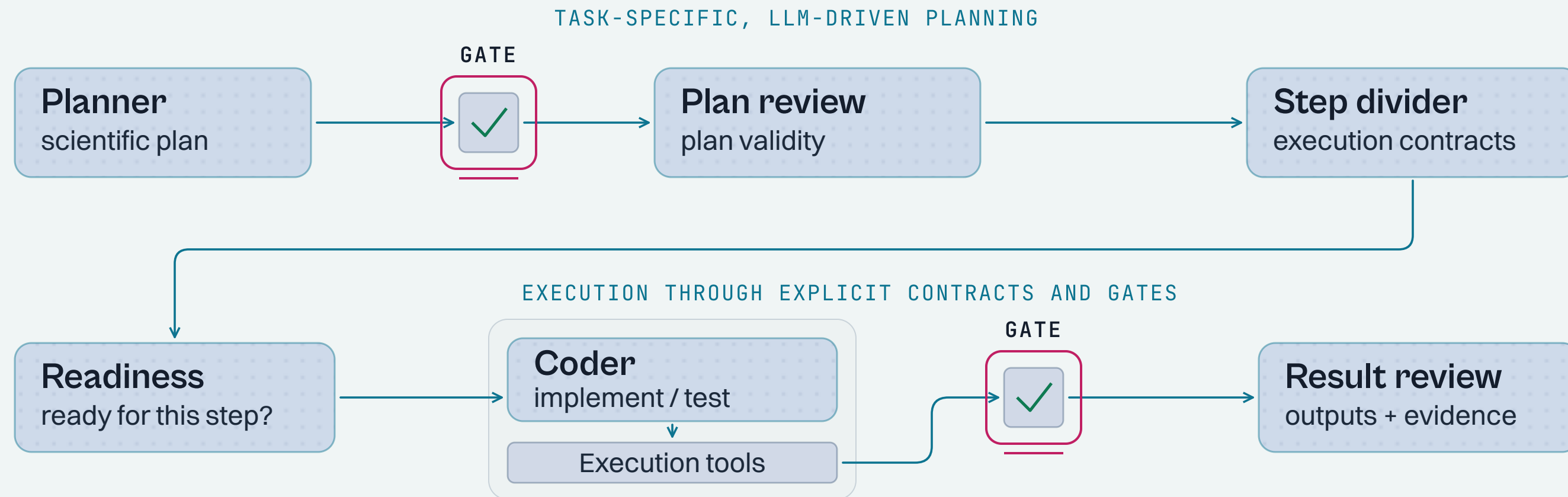
A task that runs a full simulation pipeline · one step of it, and what crosses into the next



Carrying just two of those logs forward would add ≈ 200k chars to a 300k prompt: two thirds more context that this step never reads.

Data is cleaned and cataloged; the LLM sees a unique, relevant slice.

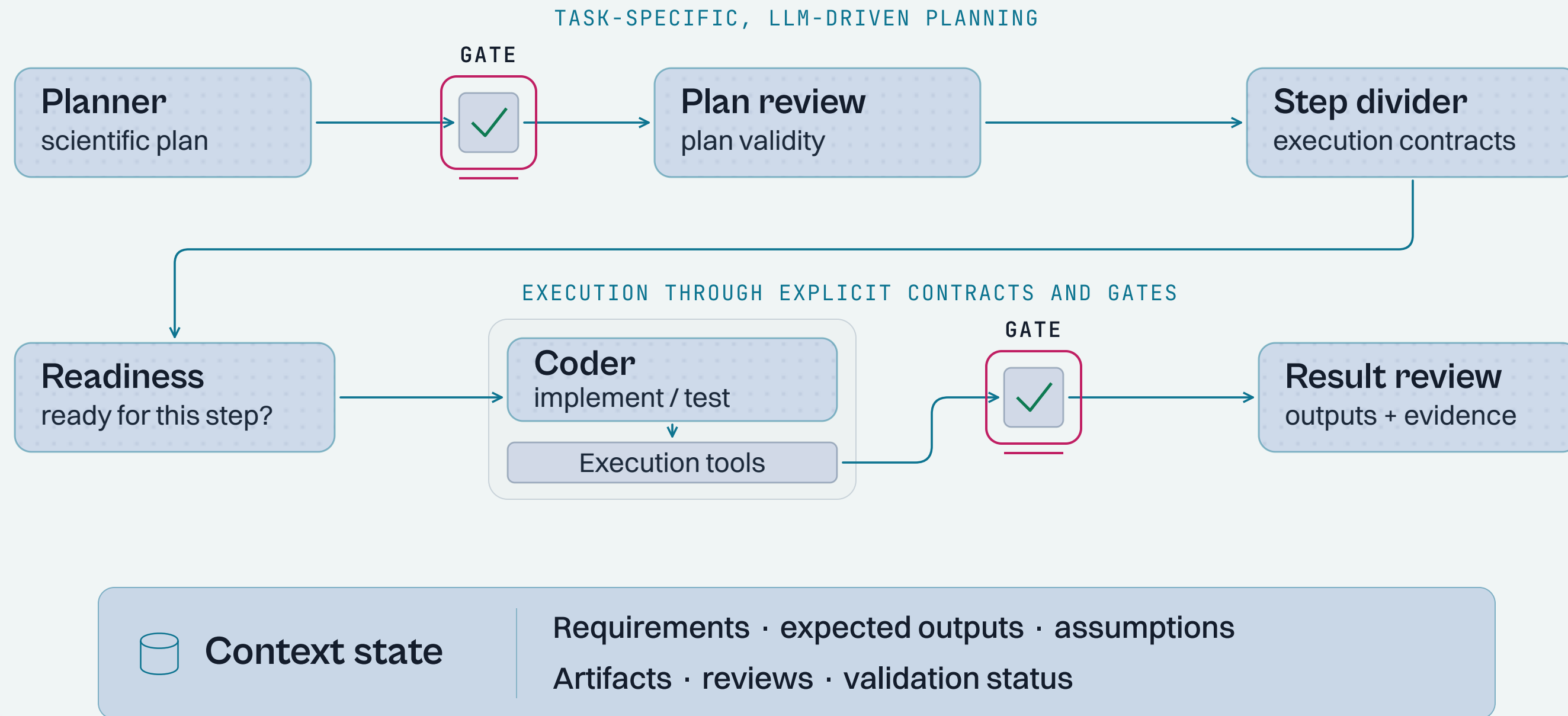
Flexible planning, structured execution



[1] Rebedea et al. (2023) · *NeMo Guardrails* · [arXiv:2310.10501](https://arxiv.org/abs/2310.10501) ↗

[2] Chitan (2026) · *ILION* · [arXiv:2603.13247](https://arxiv.org/abs/2603.13247) ↗

Flexible planning, structured execution



[1] Rebedea et al. (2023) · *NeMo Guardrails* · [arXiv:2310.10501](https://arxiv.org/abs/2310.10501) ↗

[2] Chitan (2026) · *ILION* · [arXiv:2603.13247](https://arxiv.org/abs/2603.13247) ↗

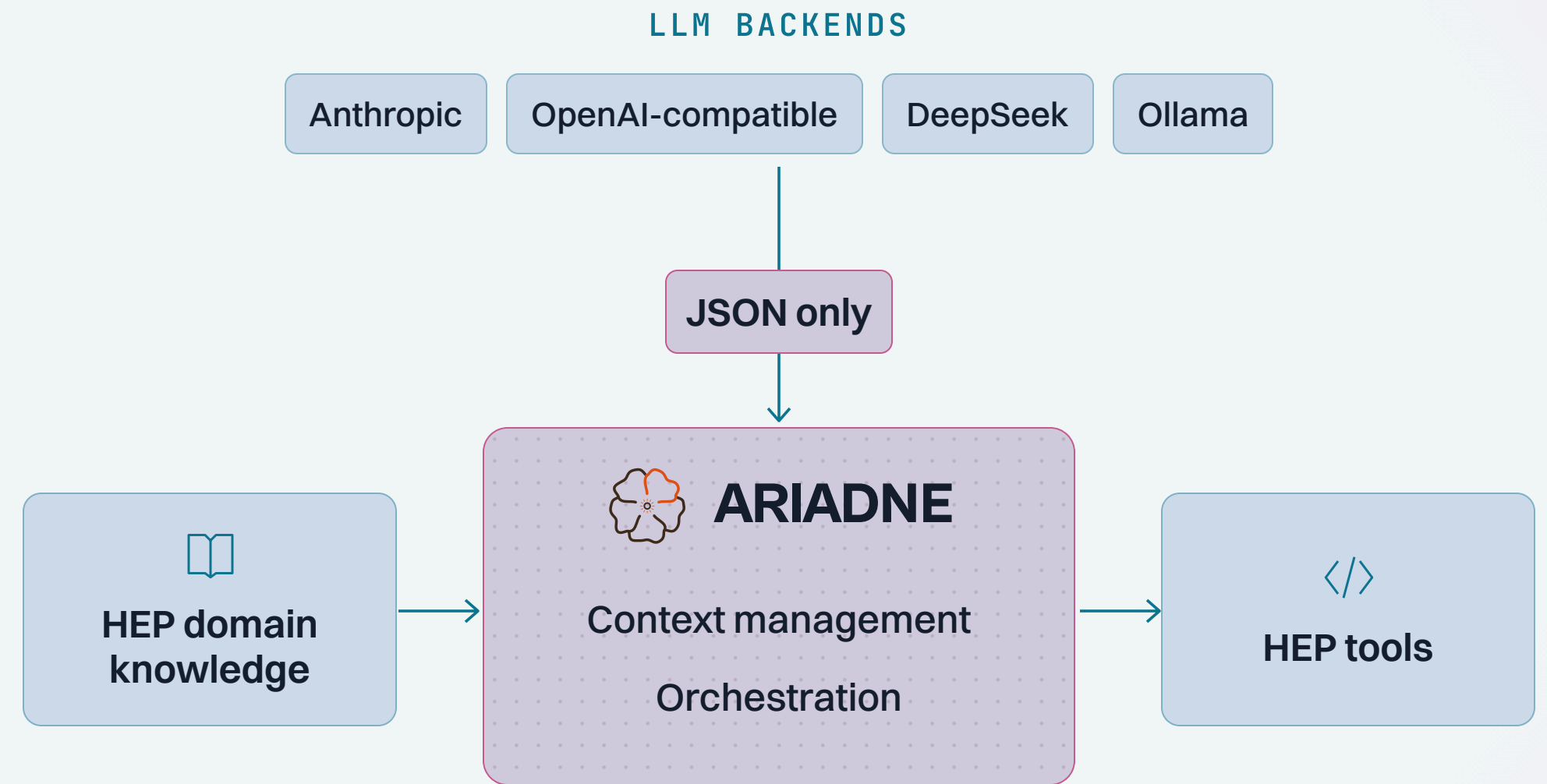
The agent chooses the path. **The harness decides when it can move on.**

The complete agent

JSON-structured output
from LLMs

Planning outputs determine downstream
gates and execution structure.

Equipped with **HEP tools and domain
knowledge**.



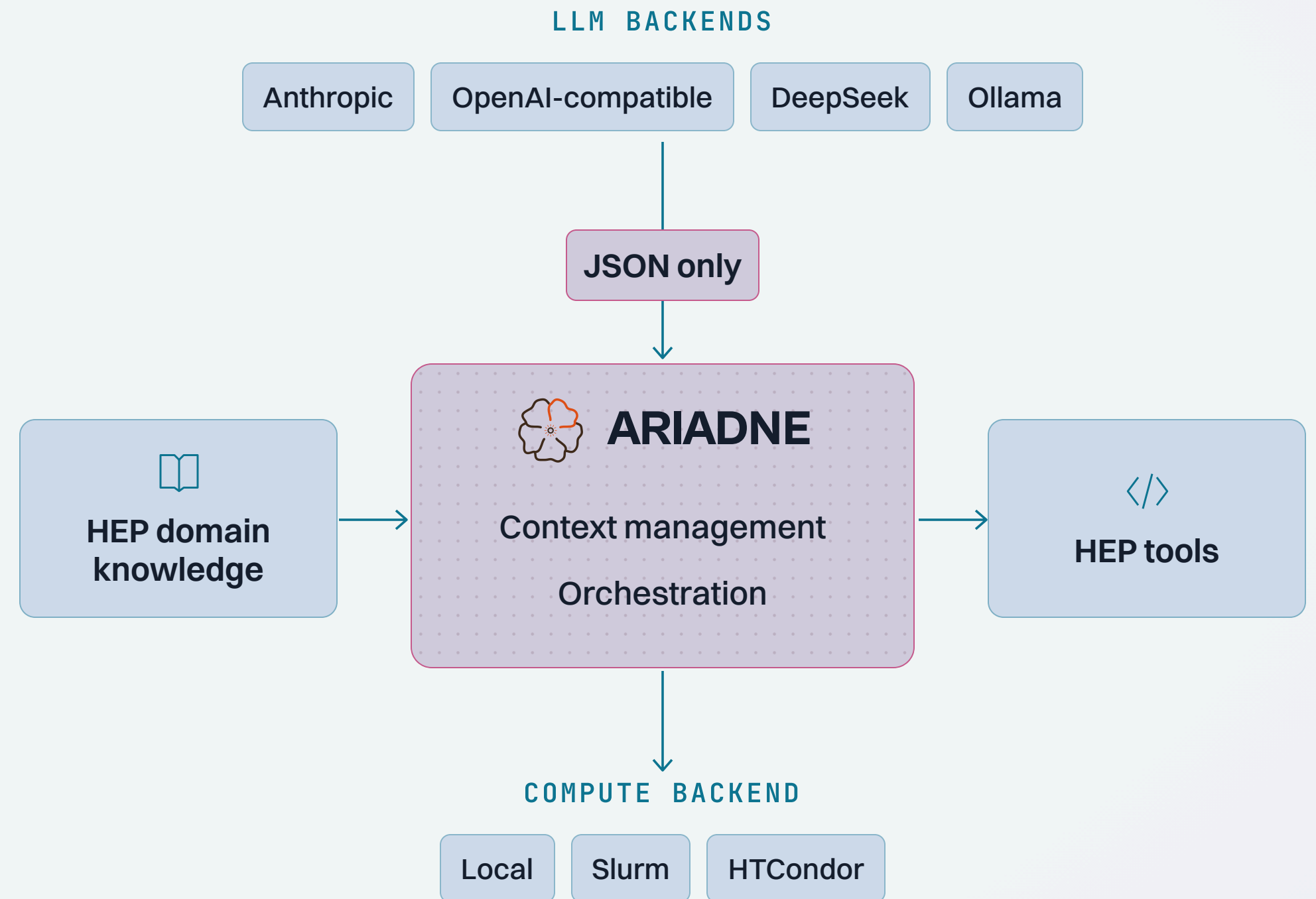
The complete agent

JSON-structured output
from LLMs

Planning outputs determine downstream
gates and execution structure.

Equipped with **HEP tools and domain
knowledge**.

Model-independent design.



Evaluation on Collider-Bench

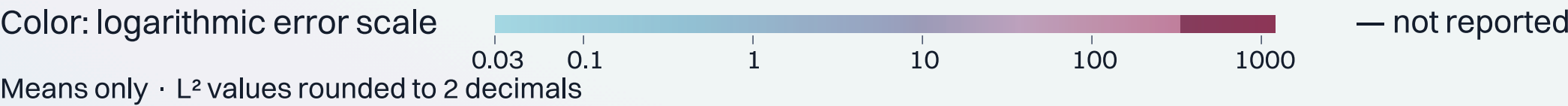
10 simulation tasks · 4 CMS searches · end-to-end simulation / recast predictions

Scored against hidden reference yields.

Faroughy et al. (2026) · Collider-Bench · [arXiv:2605.13950](#)

Relative L² error on absolute yields · lower is better

	Claude Code			Codex		ForgeCode	Physicist
	Opus 4.7	Sonnet 4.6	Haiku 4.5	GPT-5.5	GPT-5.4-mini	DeepSeek-V4	in loop
CMS-SUS-16-034							
+ TChiWZ	0.19	0.30	0.89	0.29	0.65	0.80	0.09
CMS-SUS-16-046							
T5Wg	0.51	0.61	18.00	0.13	0.50	0.99	0.05
TChiWg	0.49	0.93	0.92	0.28	0.84	0.84	0.04
CMS-SUS-16-047							
T5Wg high-HT	1.43	0.69	213.00	0.62	1.72	1.41	0.24
+ T5Wg low-HT	0.75	1.20	—	0.79	1.00	0.96	0.33
T6gg high-HT	0.64	1.27	1117.00	0.58	1.43	157.00	0.13
T6gg low-HT	11.30	1.46	278.00	0.45	0.95	0.90	0.06
CMS-SUS-16-051							
+ T2bW SRG	0.28	0.82	—	0.71	1.41	14.90	0.24
+ T2tt SRG	0.42	0.53	34.40	0.46	0.78	1.02	0.11
T2tt compressed	0.50	0.69	125.00	0.37	0.84	0.99	0.29



Evaluation on Collider-Bench

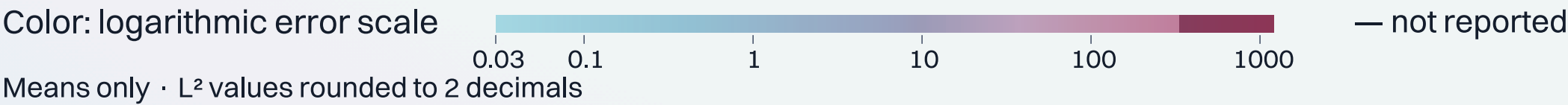
10 simulation tasks · 4 CMS searches · end-to-end simulation / recast predictions

Scored against hidden reference yields.

Faroughy et al. (2026) · Collider-Bench · [arXiv:2605.13950](#)

Relative L² error on absolute yields · lower is better

	Claude Code			Codex		ForgeCode	Physicist
	Opus 4.7	Sonnet 4.6	Haiku 4.5	GPT-5.5	GPT-5.4-mini	DeepSeek-V4	in loop
CMS-SUS-16-034							
+ TChiWZ	0.19	0.30	0.89	0.29	0.65	0.80	0.09
CMS-SUS-16-046							
T5Wg	0.51	0.61	18.00	0.13	0.50	0.99	0.05
TChiWg	0.49	0.93	0.92	0.28	0.84	0.84	0.04
CMS-SUS-16-047							
T5Wg high-HT	1.43	0.69	213.00	0.62	1.72	1.41	0.24
+ T5Wg low-HT	0.75	1.20	—	0.79	1.00	0.96	0.33
T6gg high-HT	0.64	1.27	1117.00	0.58	1.43	157.00	0.13
T6gg low-HT	11.30	1.46	278.00	0.45	0.95	0.90	0.06
CMS-SUS-16-051							
+ T2bW SRG	0.28	0.82	—	0.71	1.41	14.90	0.24
+ T2tt SRG	0.42	0.53	34.40	0.46	0.78	1.02	0.11
T2tt compressed	0.50	0.69	125.00	0.37	0.84	0.99	0.29



YIELD L²

$$d_{\text{yield}} = \sqrt{\frac{\sum_k (\hat{y}_k - y_k^*)^2}{\sum_k (y_k^*)^2}}$$

Absolute binned prediction

Collider-Bench's primary simulation fidelity metric.

SHAPE L²

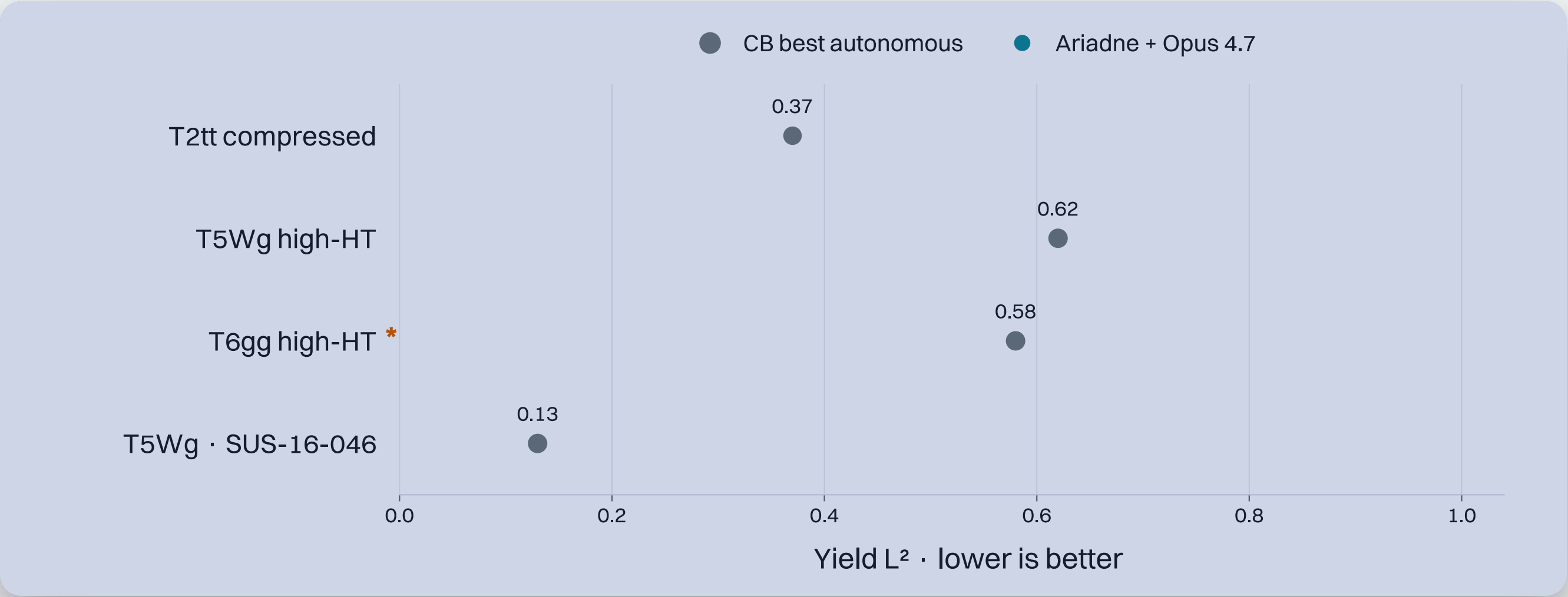
$$d_{\text{shape}} = \sqrt{\frac{\sum_k (\hat{p}_k - p_k^*)^2}{\sum_k (p_k^*)^2}}$$
$$p_k = \frac{y_k}{\sum_j y_j}$$

Distribution independent of overall normalization

0 = perfect agreement · lower is better

Ariadne vs the Collider-Bench frontier

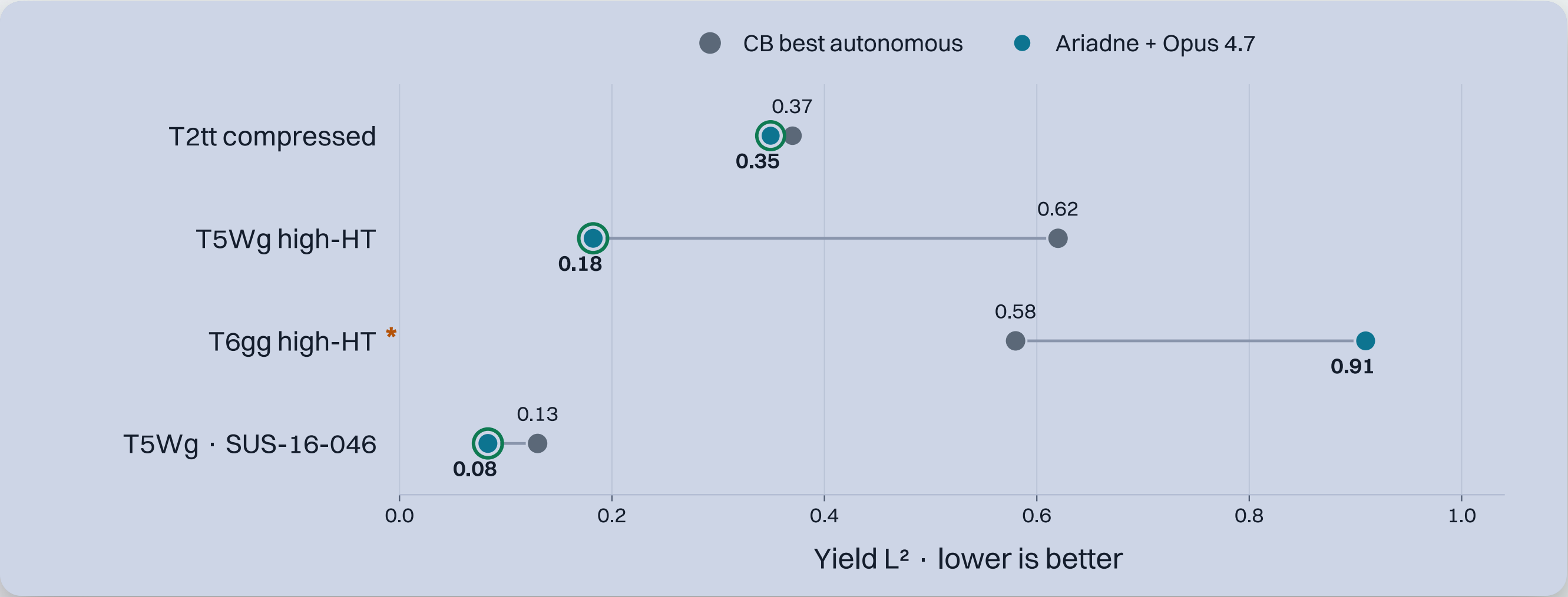
Comparing **Ariadne with Opus 4.7** against the best Collider-Bench result for each task across all tested models.



Run selection differs: the plot compares Ariadne’s best observed runs with Collider-Bench’s published means for the best-performing model on each task.

Ariadne vs the Collider-Bench frontier

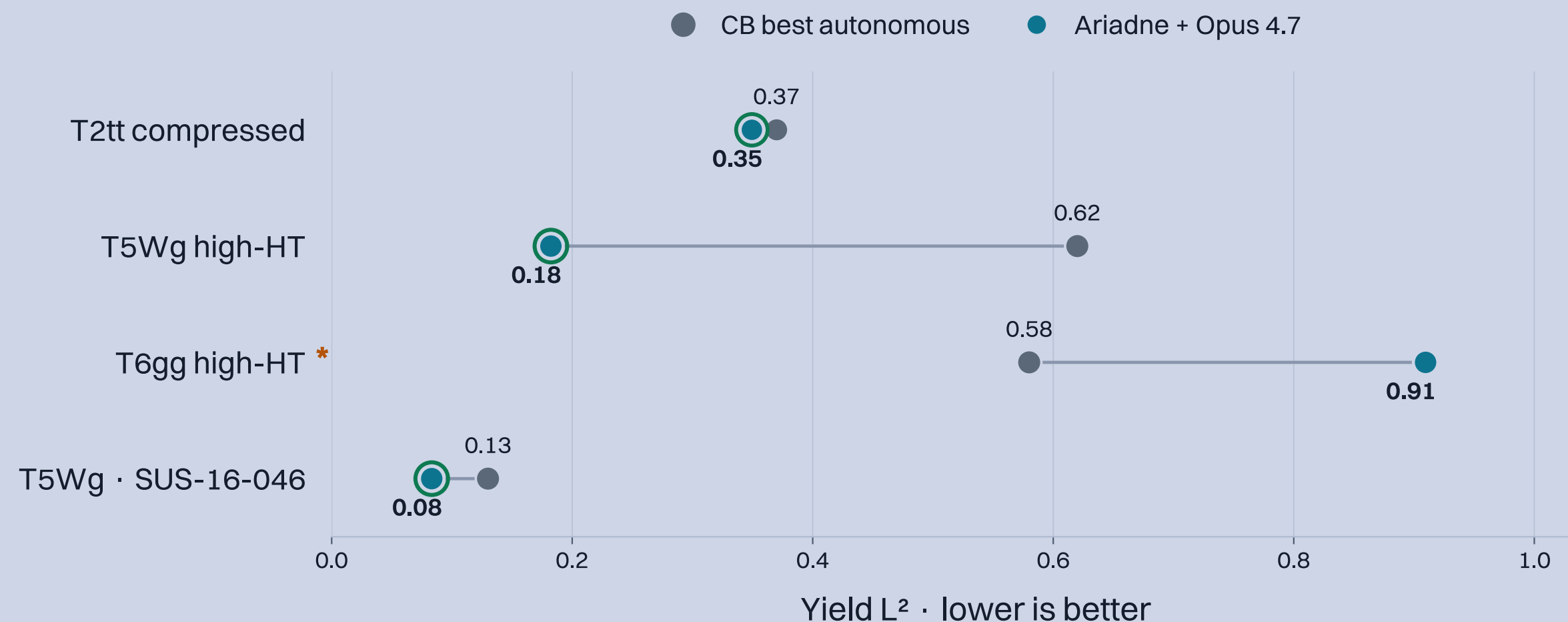
Comparing **Ariadne with Opus 4.7** against the best Collider-Bench result for each task across all tested models.



Run selection differs: the plot compares Ariadne’s best observed runs with Collider-Bench’s published means for the best-performing model on each task.

Ariadne vs the Collider-Bench frontier

Comparing **Ariadne with Opus 4.7** against the best Collider-Bench result for each task across all tested models.



* Failure mode analysis pending.

Run selection differs: the plot compares Ariadne's best observed runs with Collider-Bench's published means for the best-performing model on each task.

3 / 4

lower yield error than
the best autonomous
Collider-Bench result

T5Wg high-HT

0.51 → **0.18**

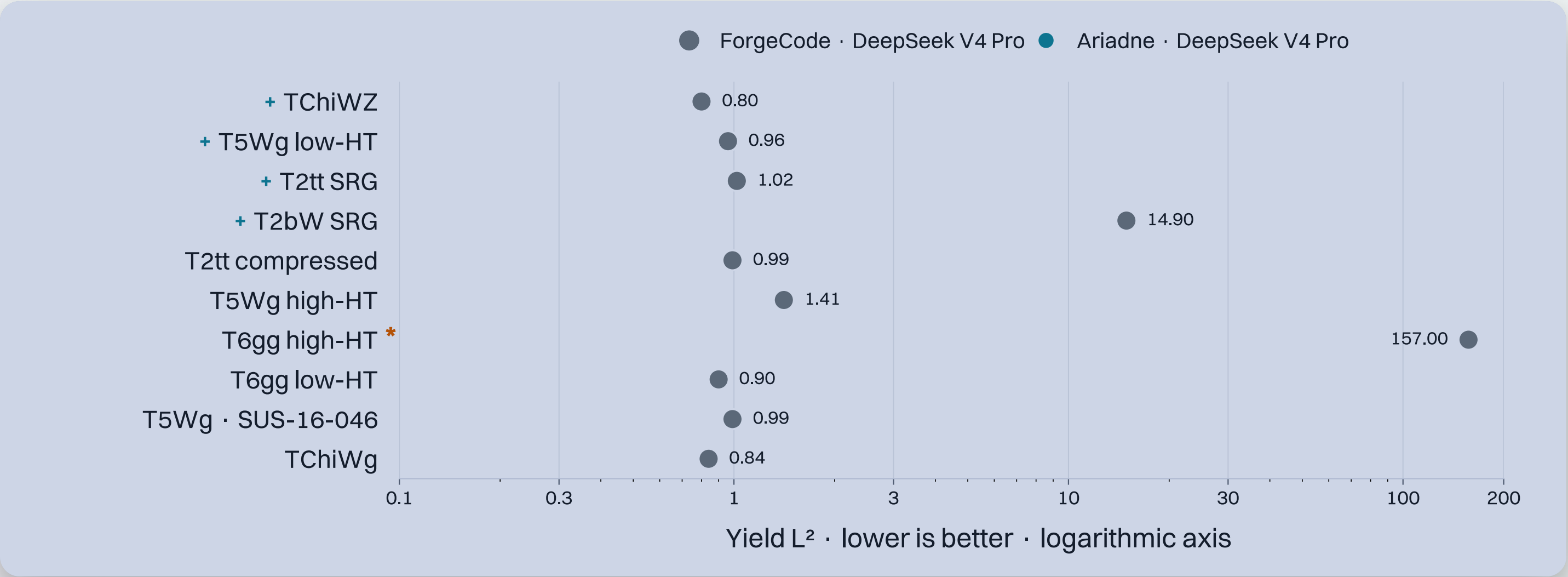
2.8× lower error

Best run vs best run
Separate from the mean-
based plot

A structured HEP harness can move the frontier beyond simply choosing the strongest benchmarked LLM.

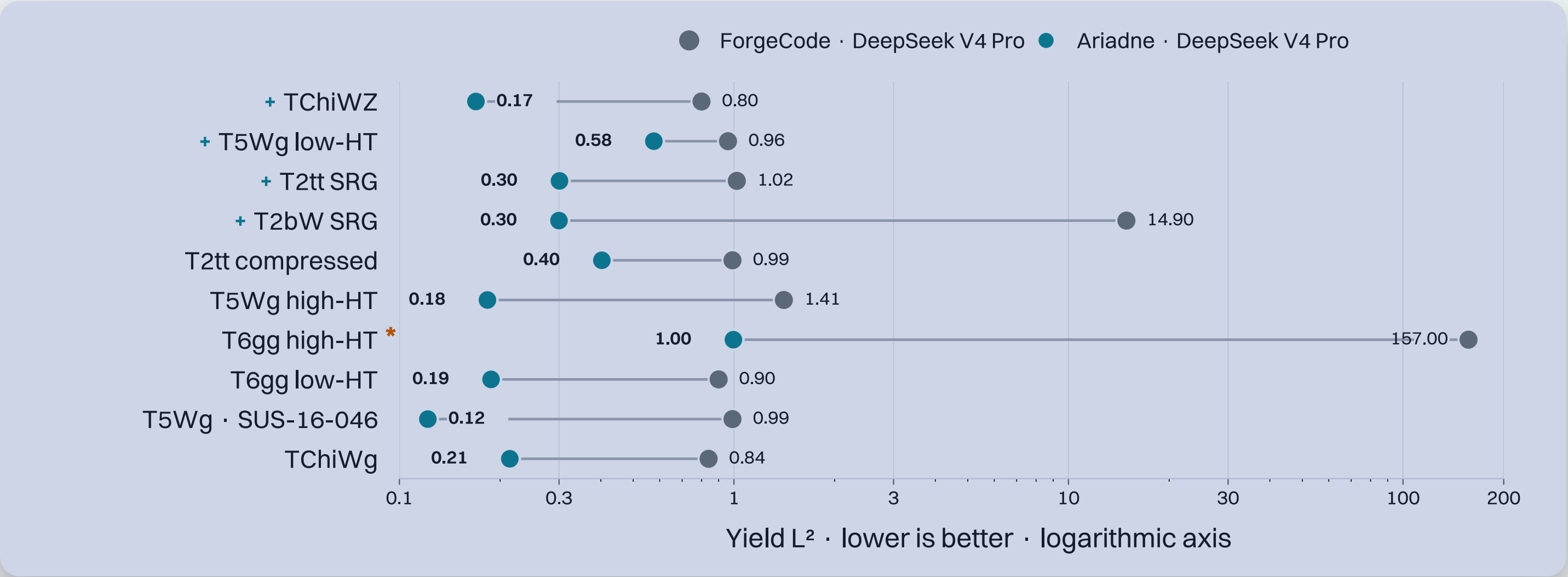
Same LLM. Different harness.

DeepSeek V4 Pro · 10 tasks Ariadne best observed runs · Collider-Bench published means



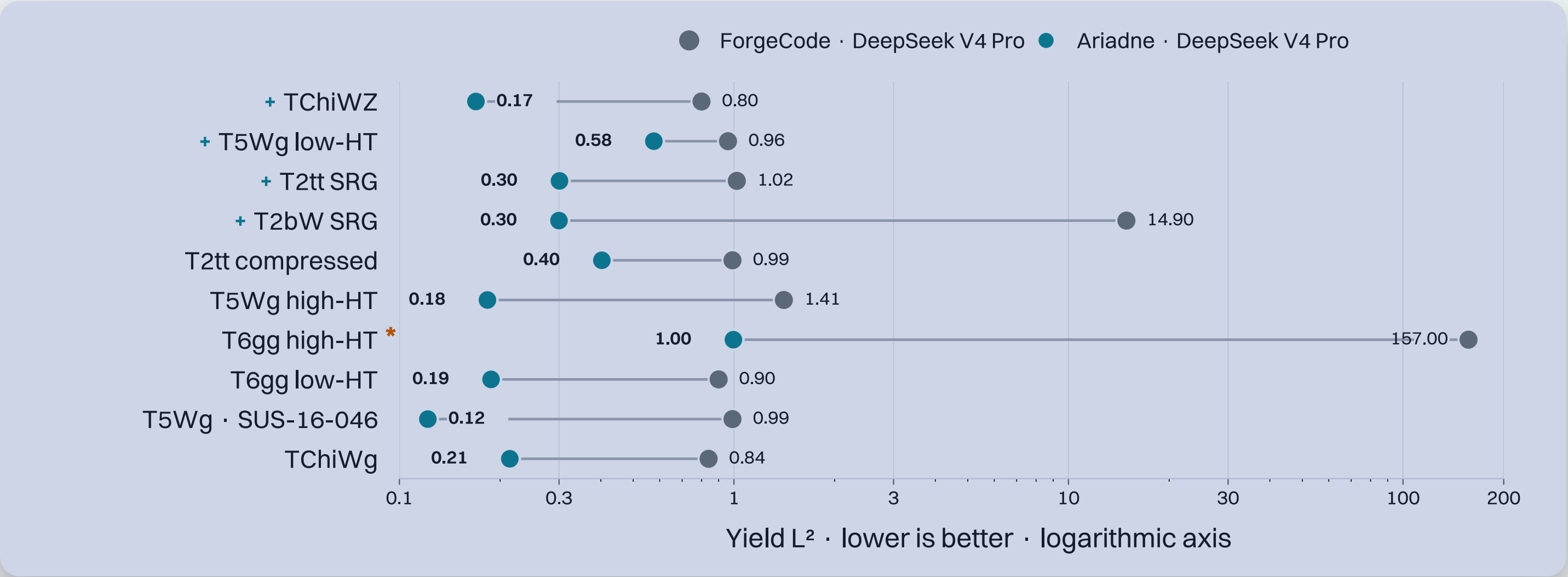
Same LLM. Different harness.

DeepSeek V4 Pro · 10 tasks Ariadne best observed runs · Collider-Bench published means



Same LLM. Different harness.

DeepSeek V4 Pro · 10 tasks Ariadne best observed runs · Collider-Bench published means



10 / 10

tasks with lower
yield L²

≈ 4.8×

median lower
error

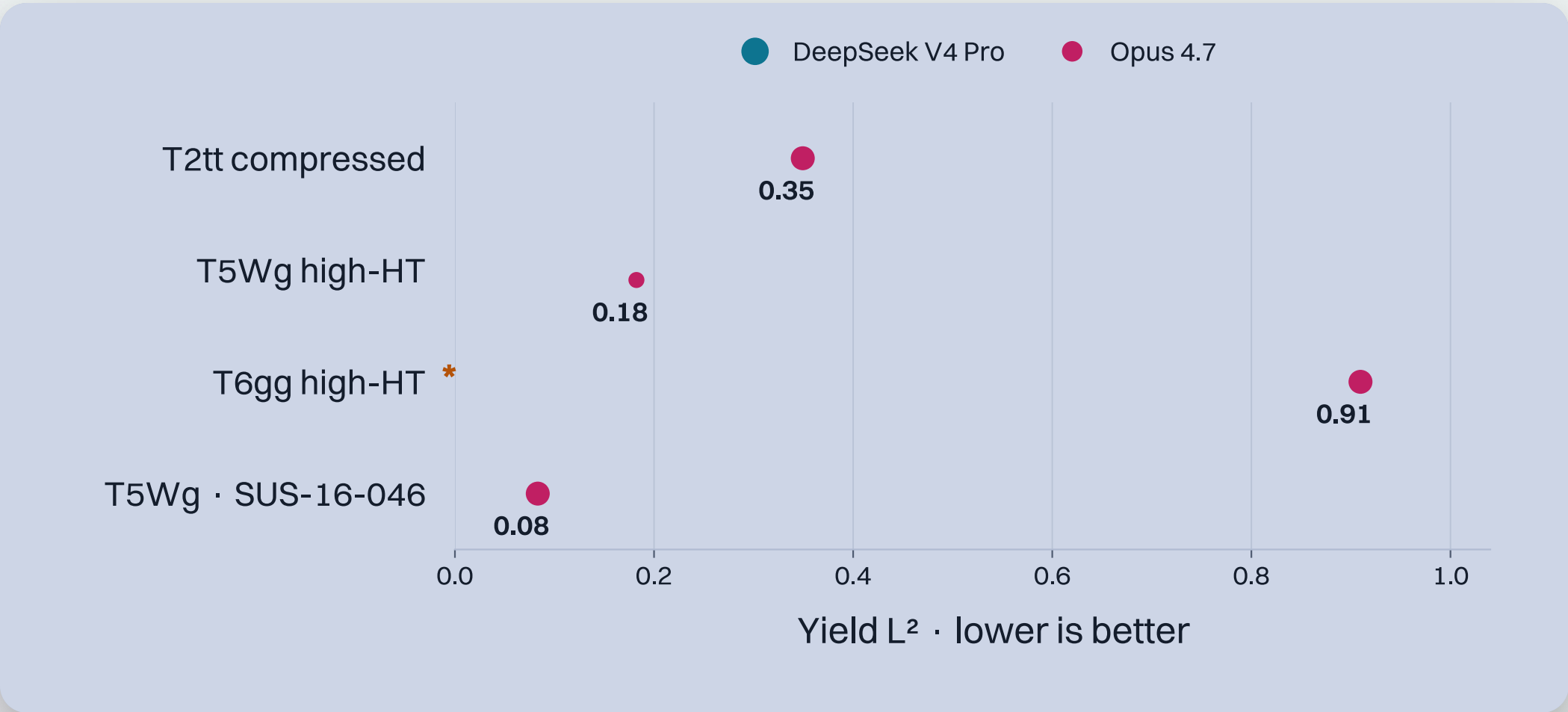
LLM held fixed

CB uncertainties omitted visually. Two extreme DeepSeek central values have uncertainties larger than their means, indicating instability. Best-run vs mean selection limits a causal harness claim.

Holding the LLM fixed helps to isolate the harness performance.

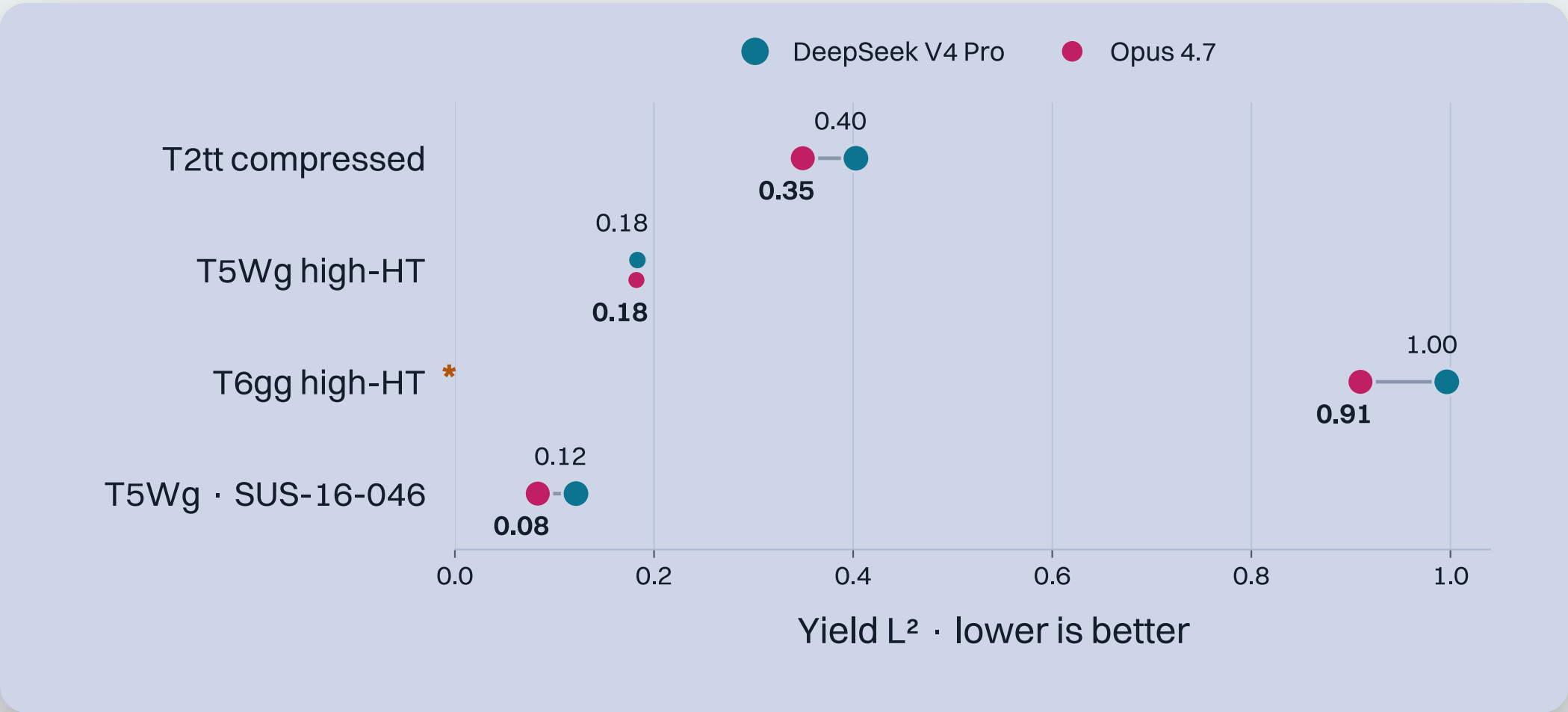
A stronger model, at a steep price

4 completed head-to-head tasks Best observed runs within Ariadne



A stronger model, at a steep price

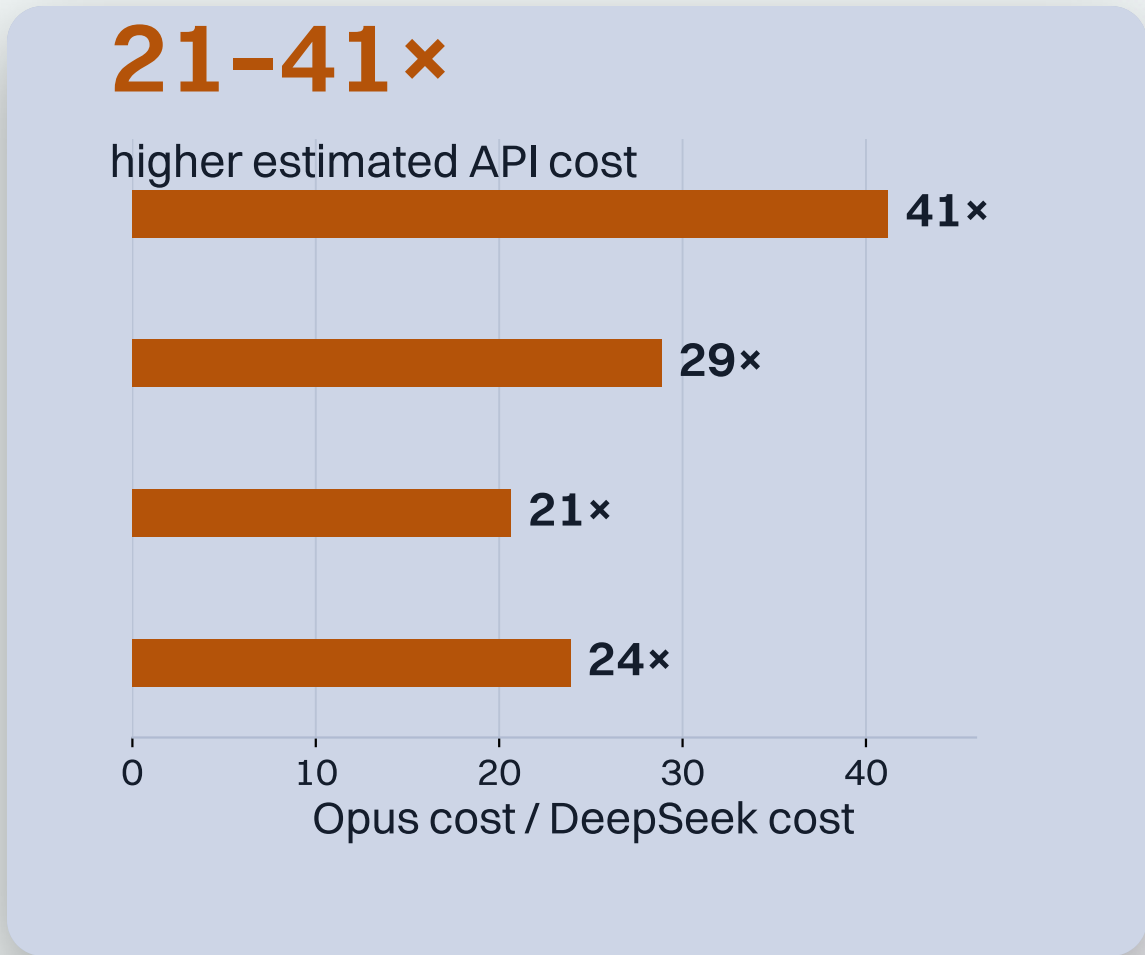
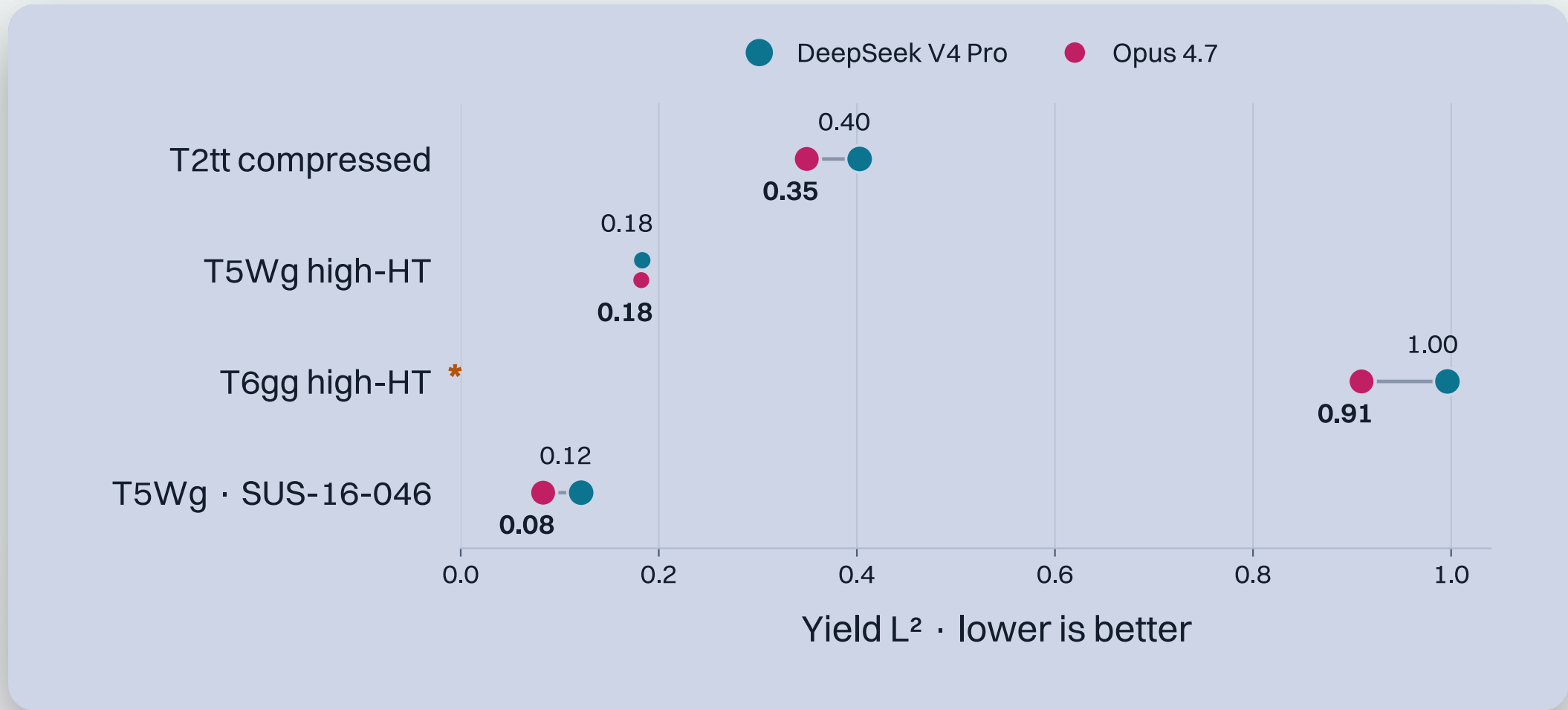
4 completed head-to-head tasks Best observed runs within Ariadne



A stronger model, at a steep price

4 completed head-to-head tasks Best observed runs within Ariadne

Opus wins 4 / 4

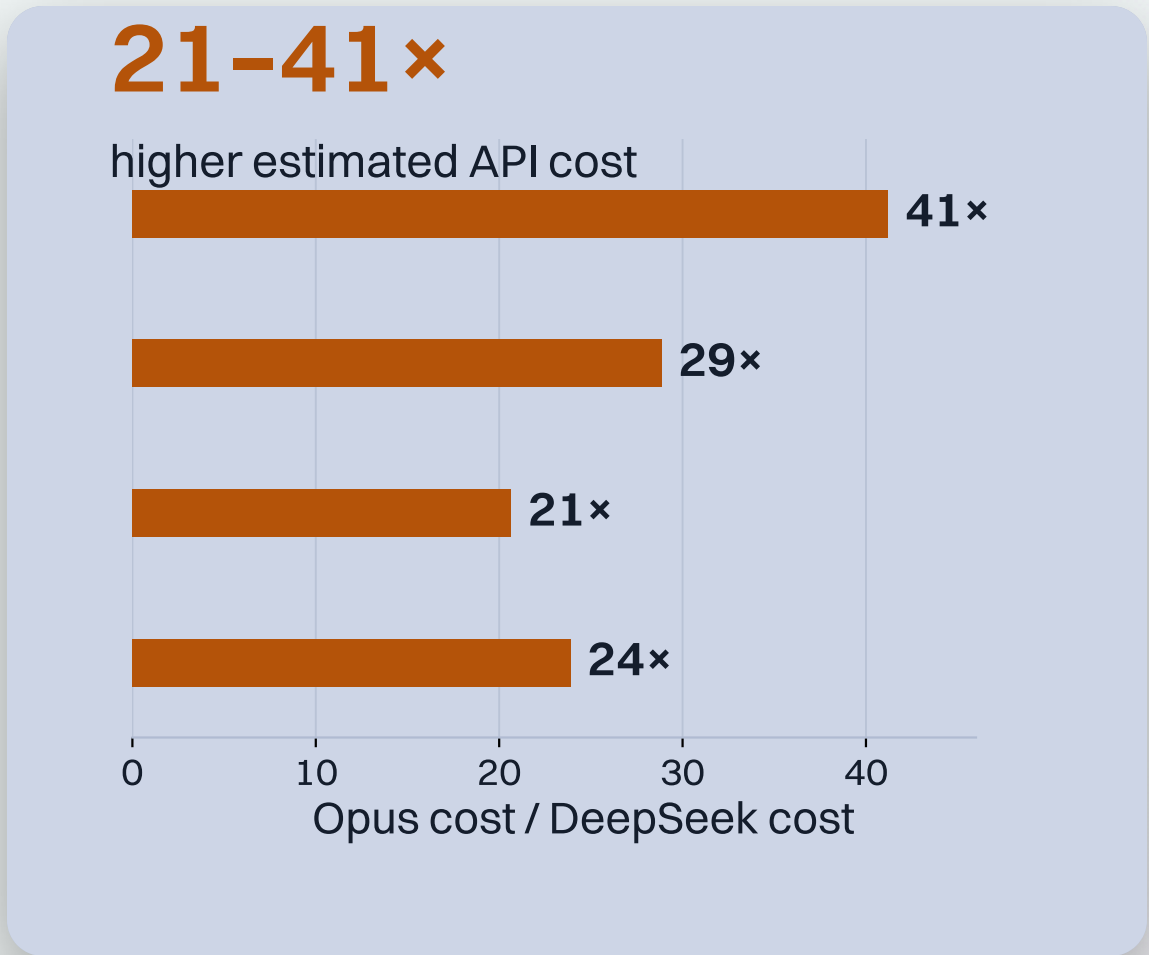
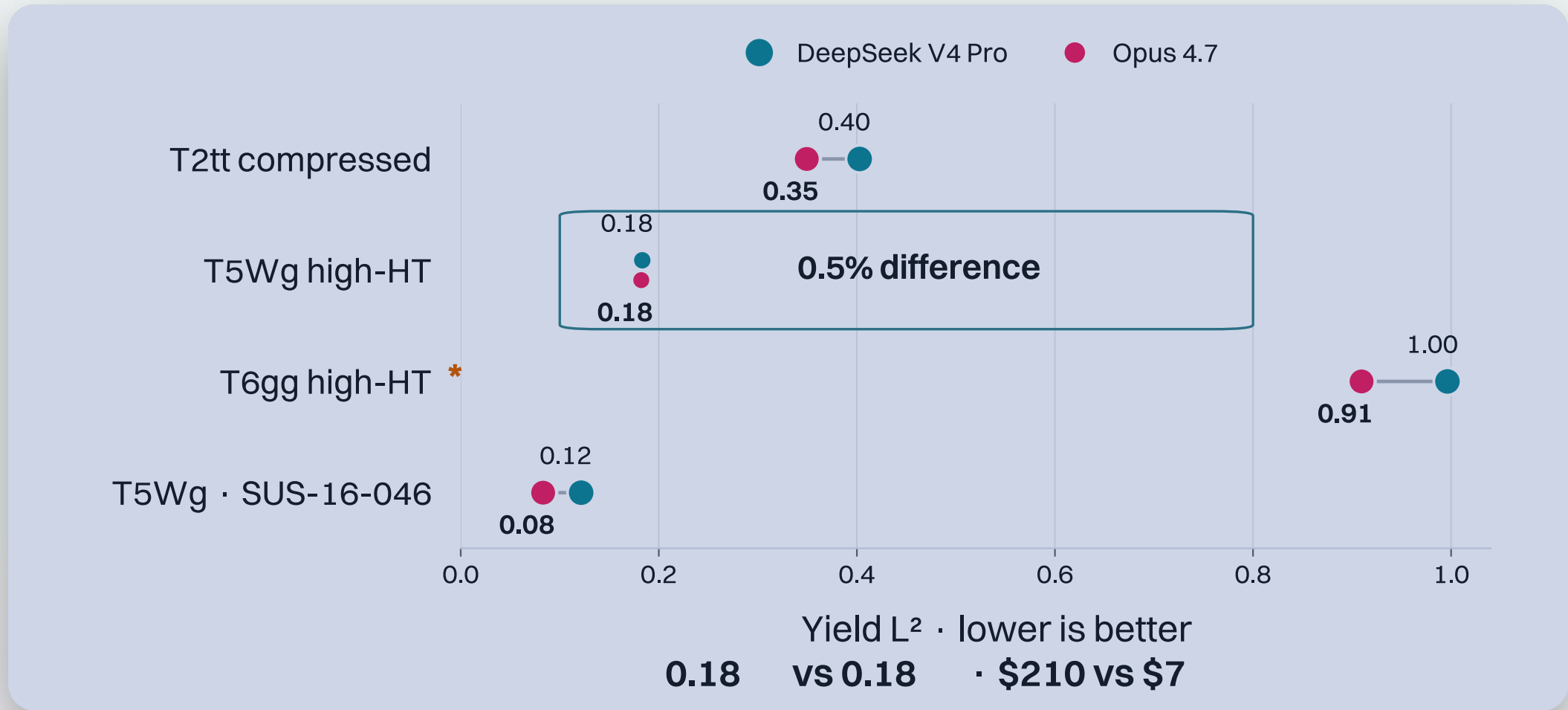


Costs are log-derived estimates. DeepSeek uses off-peak rates; treat these as lower-bound estimates, not billing records.

A stronger model, at a steep price

4 completed head-to-head tasks Best observed runs within Ariadne

Opus wins 4 / 4

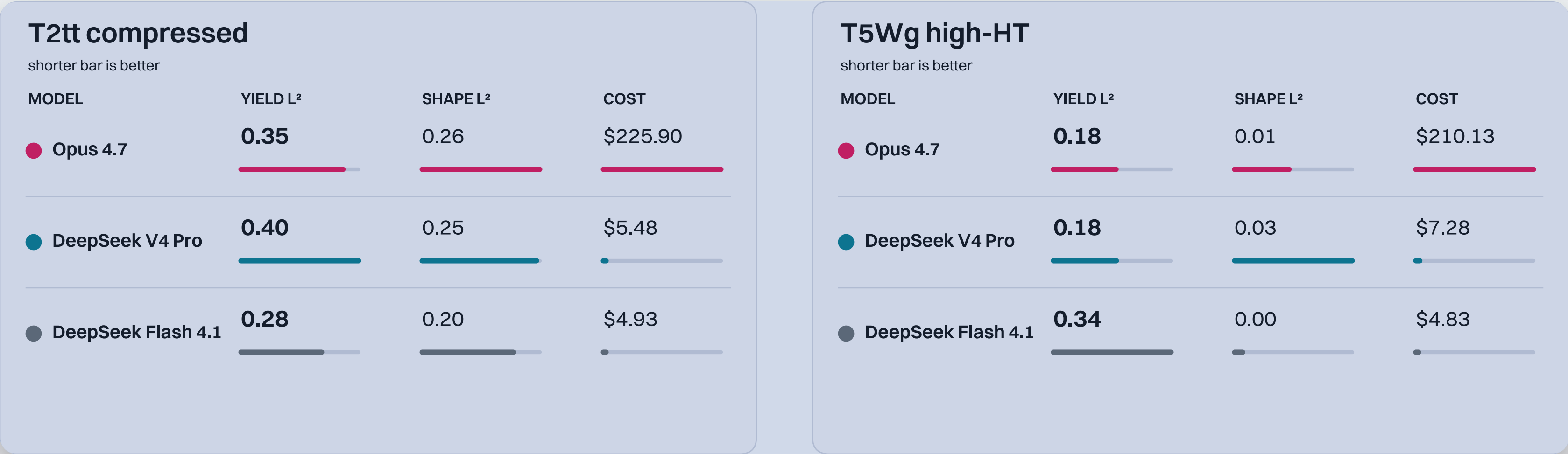


Costs are log-derived estimates. DeepSeek uses off-peak rates; treat these as lower-bound estimates, not billing records.

Better models still help, but marginal accuracy can become very expensive.

Performance comparison against an even smaller LLM

Only two tasks currently have complete three-model coverage · best observed runs

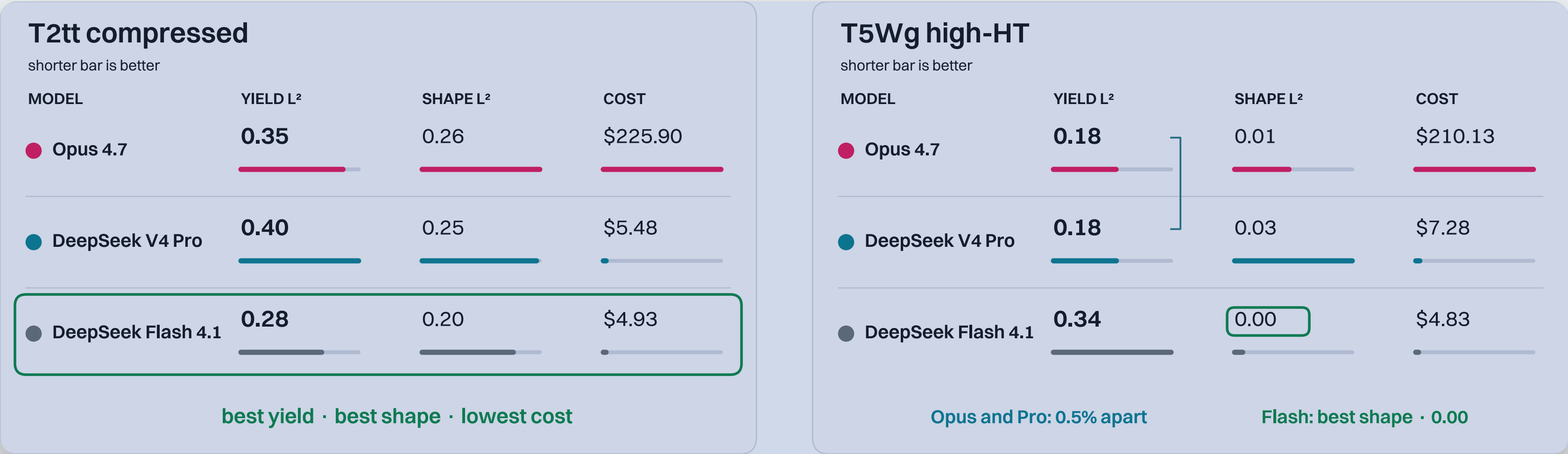


Estimated API costs use log-derived, off-peak DeepSeek rates; lower-bound estimates, not billing records.

Performance comparison against an even smaller LLM

Only two tasks currently have complete three-model coverage · best observed runs

4× B200



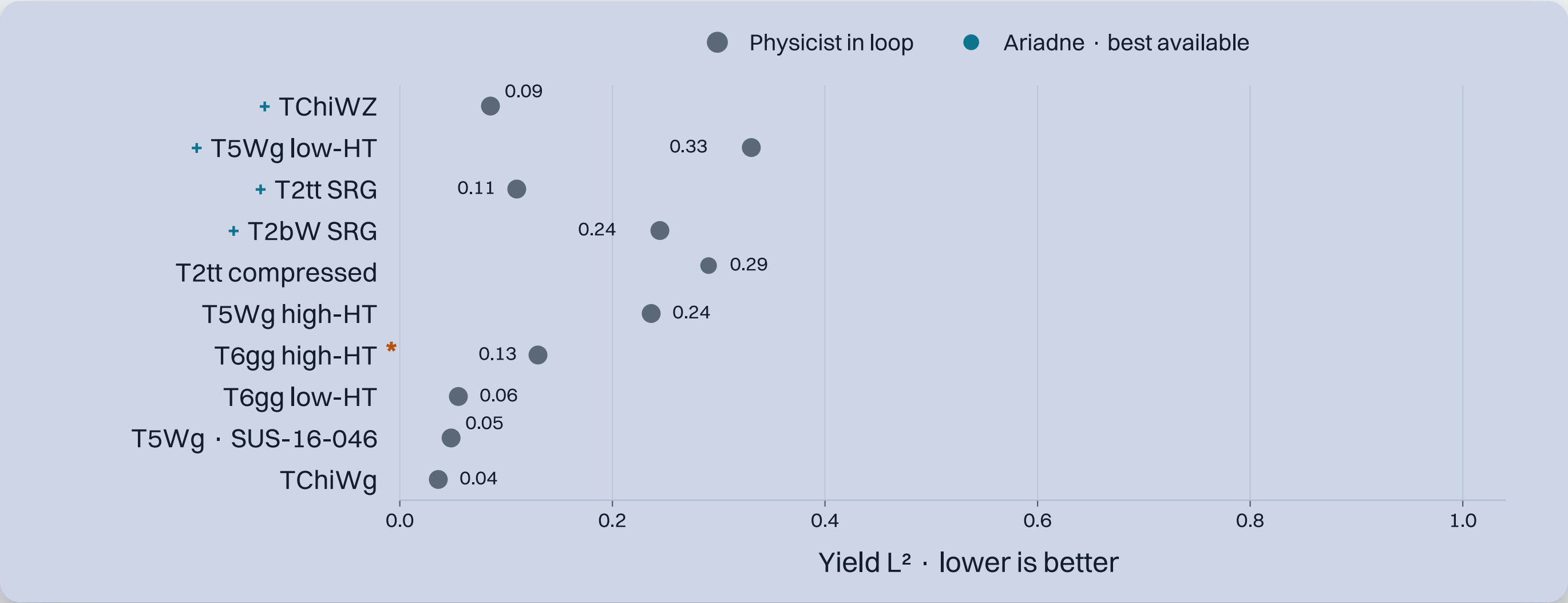
Estimated API costs use log-derived, off-peak DeepSeek rates; lower-bound estimates, not billing records.

Price does not predict the winner.

A better harness can help narrow the performance gap between low-cost and frontier LLMs.

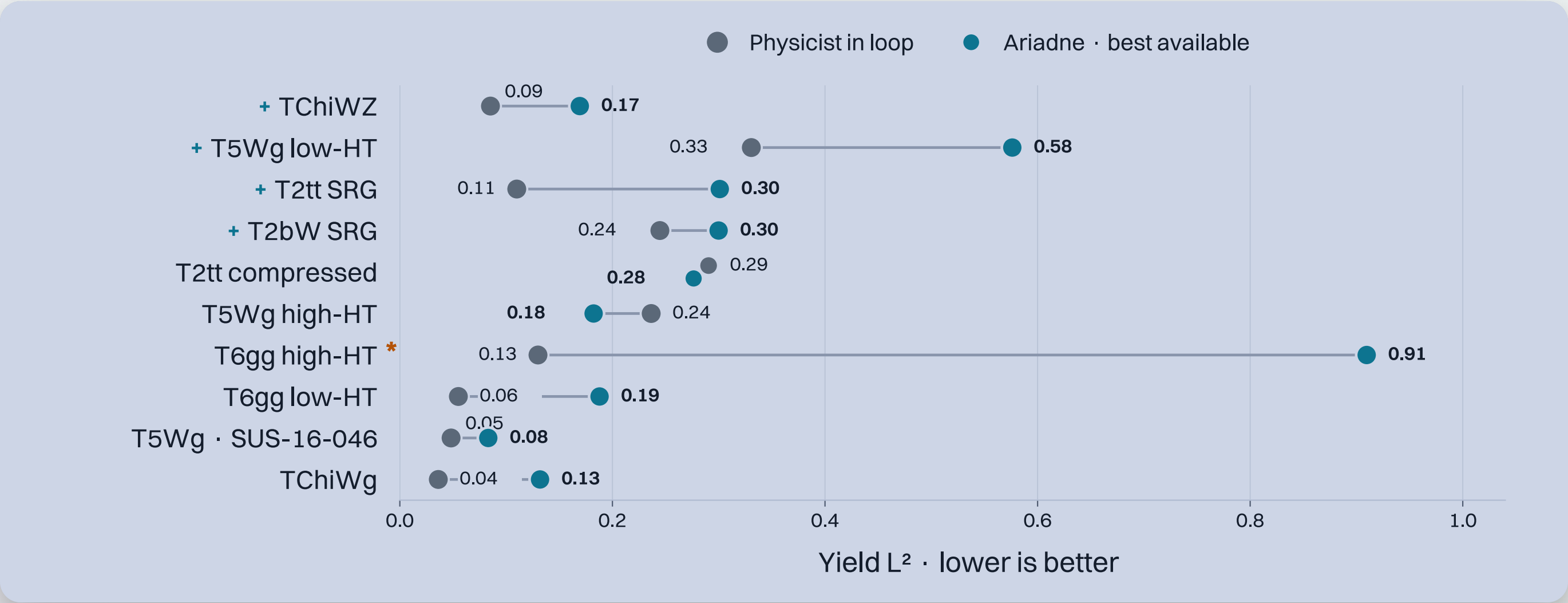
Against the physicist-in-the-loop baseline

10 tasks · best available Ariadne model per task Best observed runs vs published baseline means



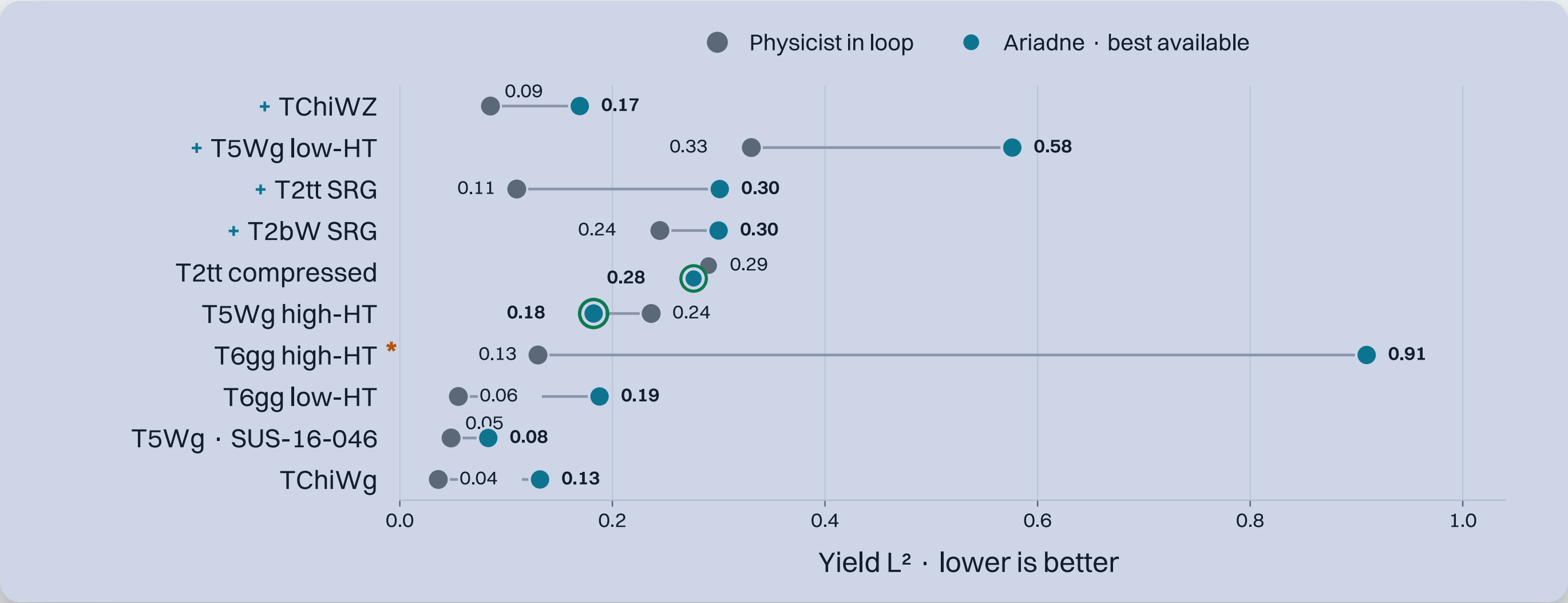
Against the physicist-in-the-loop baseline

10 tasks · best available Ariadne model per task Best observed runs vs published baseline means



Against the physicist-in-the-loop baseline

10 tasks · best available Ariadne model per task Best observed runs vs published baseline means



2 / 10

lower central yield L² than the human-assisted baseline

6 / 10

within 2× or better

Central values shown without error bars. Several close comparisons may change interpretation once uncertainties are included.

Not human-level autonomy yet. The remaining gap is already strongly task-dependent.

Summary and outlook

Summary

- **An agent built for HEP:** combines established HEP tools and domain guidance with new approaches to context management and workflow control.
- **First in HEP:** a dedicated context manager built around structured LLM outputs, preserving relevant information across long scientific workflows.
- **Flexible planning, structured execution:** uses context state to enforce deterministic and plan-defined gates throughout the workflow.

Summary and outlook

Summary

- **An agent built for HEP:** combines established HEP tools and domain guidance with new approaches to context management and workflow control.
- **First in HEP:** a dedicated context manager built around structured LLM outputs, preserving relevant information across long scientific workflows.
- **Flexible planning, structured execution:** uses context state to enforce deterministic and plan-defined gates throughout the workflow.

Outlook

- **Locally deployed models:** extend Ariadne to more compact, self-hosted LLMs for even better cost efficiency.
- **Role-specific LLMs:** test different models for planning, coding, research, and review.
- **Beyond recasting:** evaluate broader HEP tasks and benchmarks to test its general-purpose scope.
- **Beyond HEP:** adapt the guidance and tools to other scientific fields while retaining the core architecture.

ARIADNE · ML4Jets 2026 · Vienna

Thank you.

Questions?

D. Faroughy · S. Palacios Schweitzer · I. Pang

D. Shih · AU

Rutgers University

aman.u@rutgers.edu

ML4Jets 2026 • Vienna

Backup slides

Why another HEP agent?

What we already have



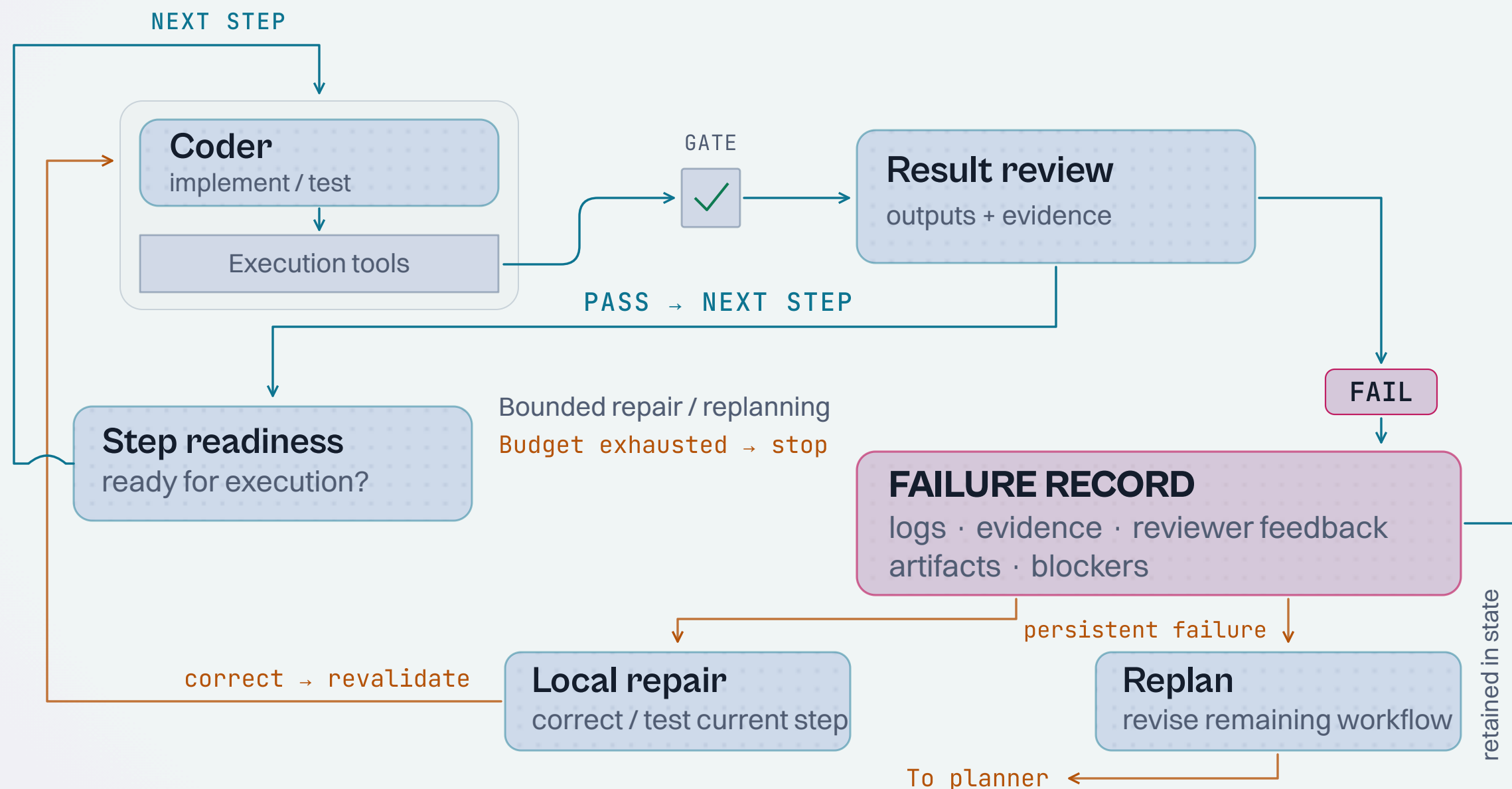
What we aim to improve

Consistently, across the full workflow.

-  Long-horizon coherence
-  Cost-efficient execution
-  Reproducible and reliable
-  LLM-agnostic

The question is no longer whether agents can help HEP,
but how to make them **reliable over an entire scientific workflow.**

Failure is part of the workflow



A bare retry

error → “try again”

Ariadne

record → diagnose → repair / replan
→ revalidate

ILLUSTRATIVE FAILURE

Process / model configuration

Execution

MG5 fails / inconsistent output

Review

Required decay missing from LHE

Evidence retained

Command + stderr · process card

model choice · source finding

Repair

Modify process / card and rerun

Replan if the approach is wrong

Revise the remaining workflow



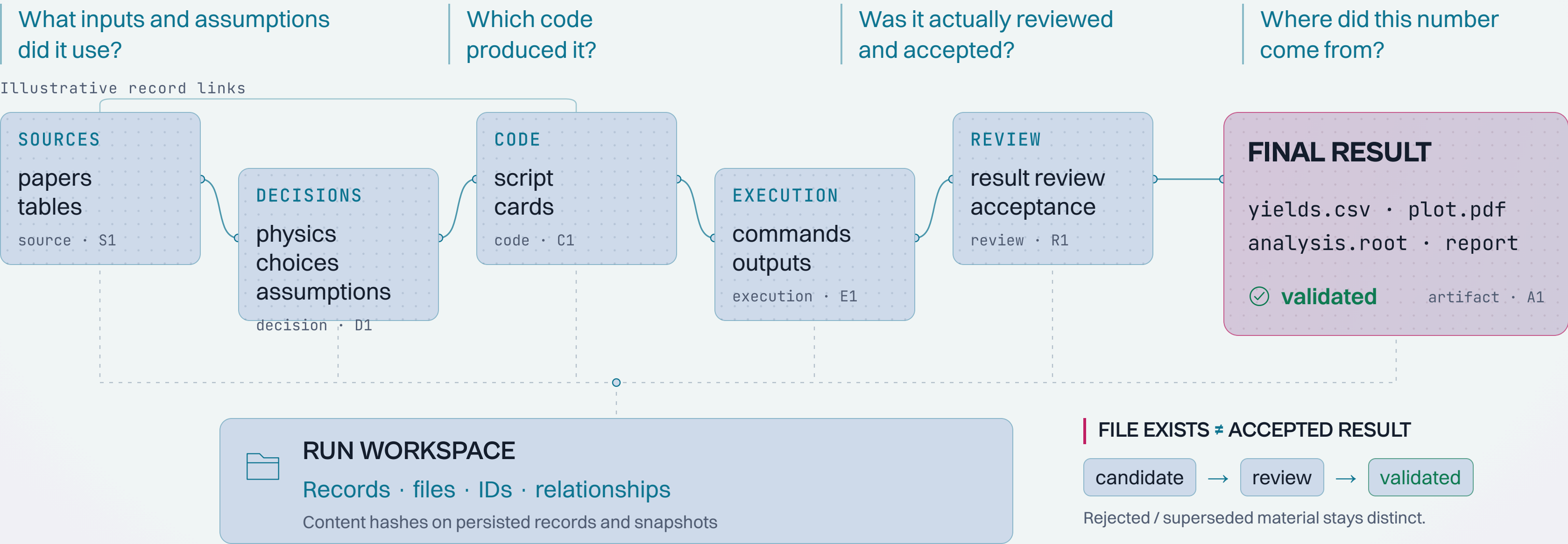
CONTEXT STATE

The next attempt inherits the evidence.

The goal is not to prevent every mistake, but to **make mistakes recoverable**.

Every result leaves a trail

Can I answer these questions after the run?



Every accepted result should have a traceable path back to its evidence.

Example task, end to end

TASK INPUT • CONDENSED

CMS-SUS-16-047

T5Wg_1600_100

$\sqrt{s} = 13 \text{ TeV} \cdot 35.9 \text{ fb}^{-1}$

Task:

Implement the CMS search and predict the binned signal yield in p_T^{miss} for the high- H_T^Y signal region.

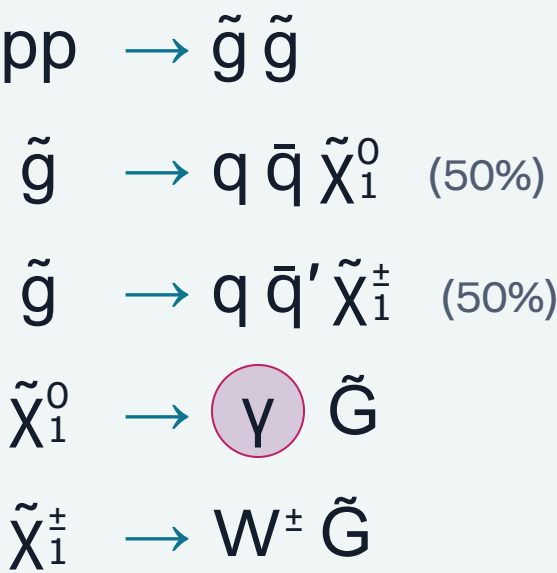
Generate events → determine cuts from the paper

→ apply analysis → fill the provided YAML.

Do not copy the signal yields from the paper or HEPData.

- Provided
- CMS paper
 - Photon-efficiency map
 - Empty result template
- Required outputs
- Analysis code
 - Simulation / cards
 - Selected events
 - Filled result
 - Report

PRODUCTION & DECAY CHANNELS



T5Wg_1600_100

$m_{\tilde{g}} = 1600 \text{ GeV}$

$m_{\tilde{\chi}} = 100 \text{ GeV}$

Common neutralino / chargino mass



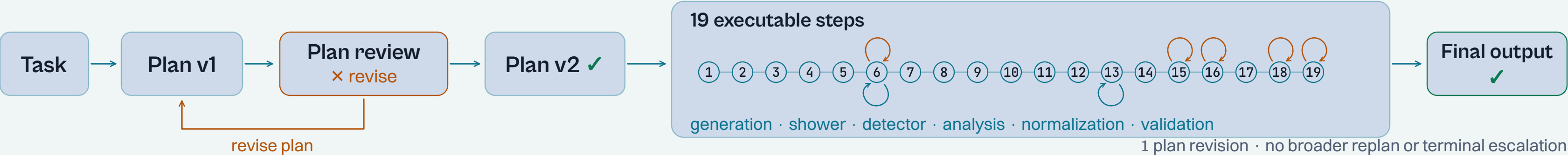
From this specification, Ariadne had to **reconstruct the analysis workflow autonomously**.

The run was **not** a straight line

19/19 steps completed

87 graph iterations **114** command batches

26 submissions → 19 accepted · 5 coding retries · 2 requested tests



PLAN REVIEW

Caught before simulation

“embedded `<slha>` block ... takes precedence over `SLHA:file`”

Possible consequence:
intended BSM decays never happen.

RESULT REVIEW

Caught after implementation

“uses the generator-level GenJet collection for H_T^Y ... whereas CMS ... defines jets as reconstructed...”

Wrong selection, biased yields.
Source-grounded review sent it back.

FINAL VALIDATION

Caught in provenance

“that file is absent from the workspace ... so the statement is false”

Methodology claims were rejected; evidence of pipeline-produced yields was also required.

Step-result review decisions

■ Accepted 19 ■ Coding retry 5 ■ Requested tests 2



Autonomy came from **recovery**, not from getting every decision right the first time.

What testing has taught us

Long workflows now complete reliably. Testing exposes recurring bug classes, which we harden structurally; their frequency has fallen considerably.

01 CONTEXT MANAGEMENT

Store the right information · route it to the right node

- Important information never entered persistent state
- Information exists, but is not rendered where needed

testing → routing and handoff fixes

02 BSM PHYSICS

Models · SLHA · decays · non-default UFOs

- Choosing or constructing the correct signal model
- SLHA / parameter-card handling
- Less standard BSM topologies

testing → stronger physics guidance + validation

* T6gg high-HT · recent unresolved generation issue

03 DEBUGGING & RECOVERY

When to repair · when to continue · when to replan

- Do not terminate a run at the first modest discrepancy
- Give the agent room to diagnose and recover
- Escalate only when evidence indicates the approach itself is wrong

testing → better recovery policy

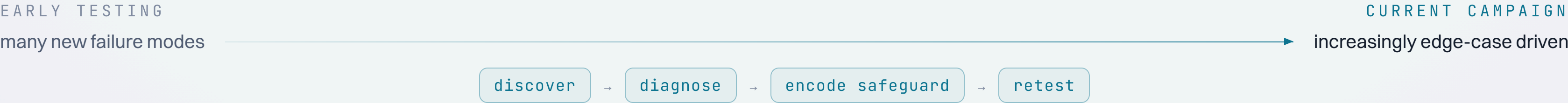
04 TOOLS & INFRASTRUCTURE

Wrappers · configuration · distributed execution

- External software quirks
- Silent configuration behavior
- Cluster/runtime issues

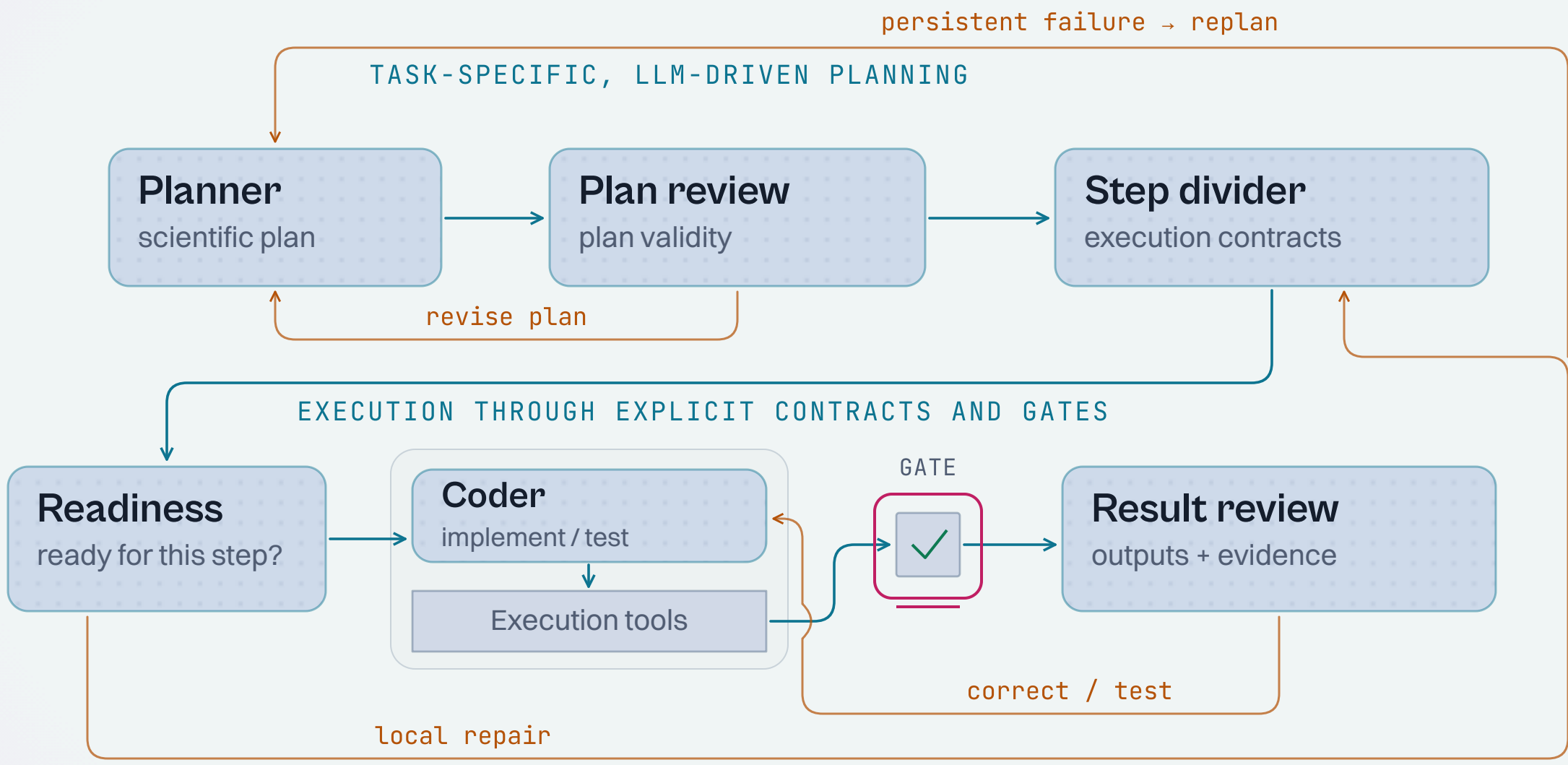
file exists ≠ tool used it correctly

testing → stronger pre-checks and audits



Testing has shifted from finding basic failures to hardening edge cases.

Flexible planning, structured execution



What counts as “done”?

Hard-coded check

required output / schema / command status

Plan-defined check

expected deliverable → enforced later

EXAMPLE CONTRACT

Plan	generate events
Expected	LHE + audit record
Gate	missing deliverable → step cannot pass

Scientific review

LLM judges evidence / physics

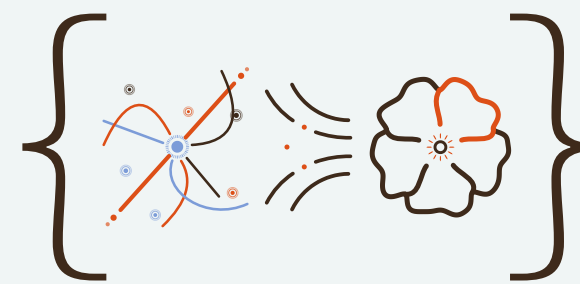


Context state

contracts carried across steps

requirements · expected outputs · assumptions · artifacts · reviews · validation status

The agent chooses the path. **The harness decides when it can move on.**

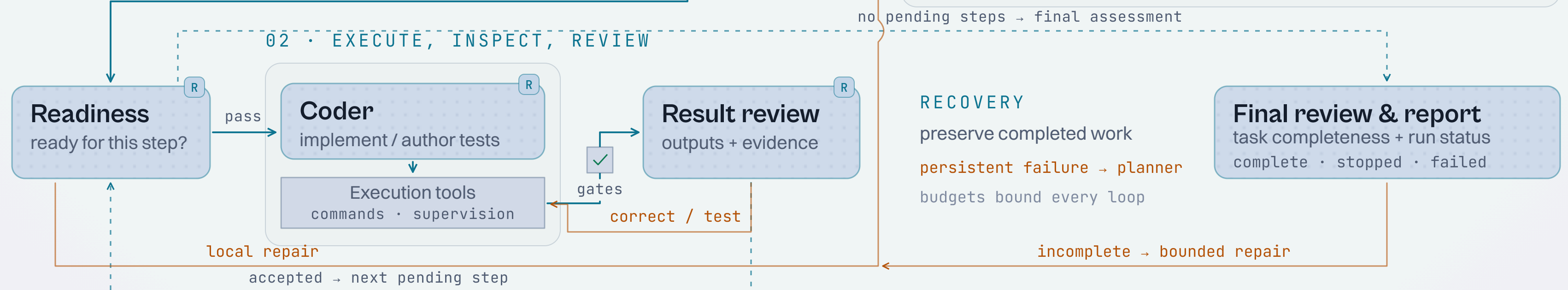
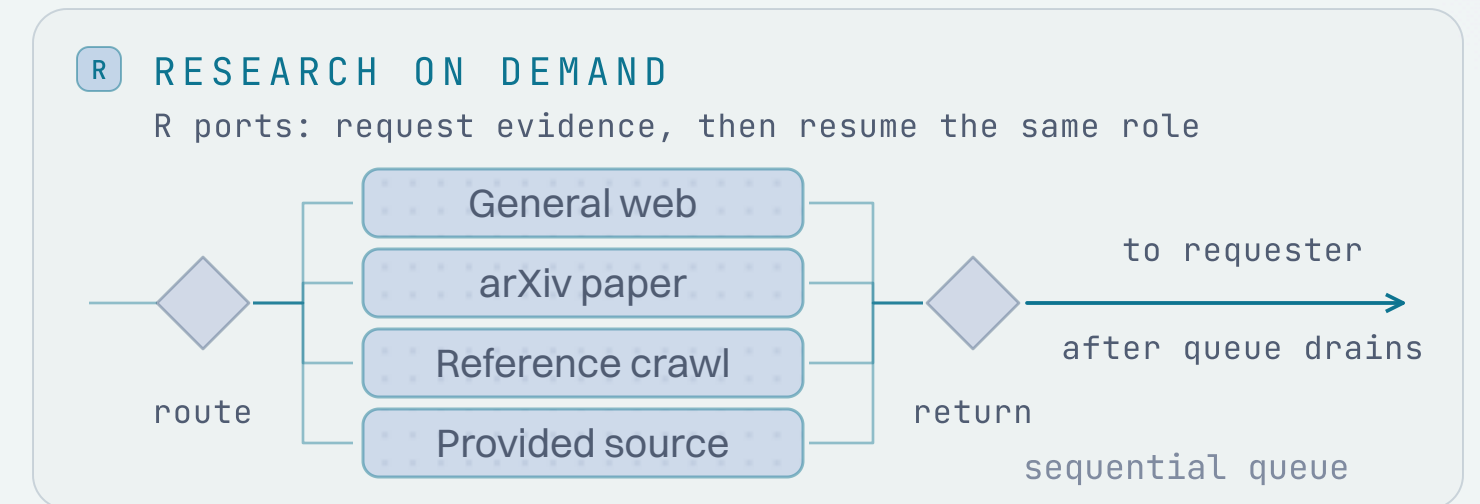
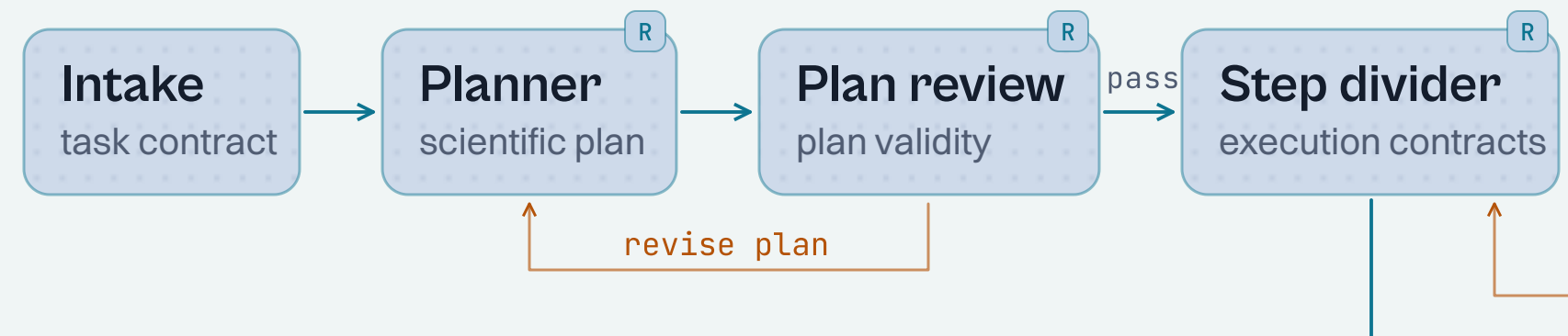


Ariadne

□ Node → Repair / retry → Workflow progression

01 • FORM THE SCIENTIFIC PLAN

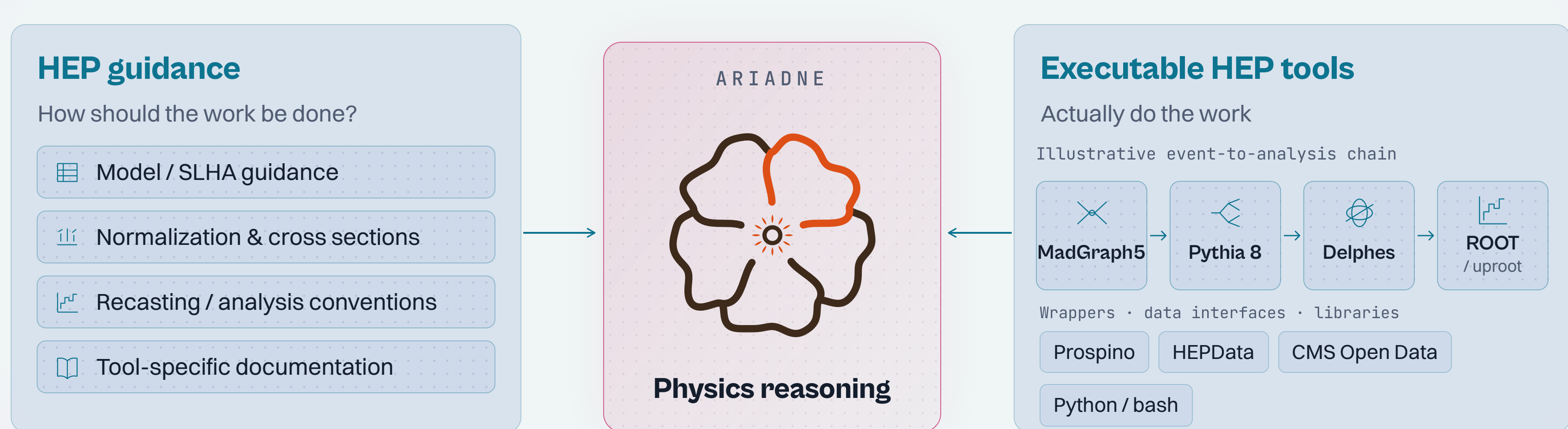
task + assets → workspace preparation



**Context state**
shared across the workflow

task · plan · steps · notes · research · workspace · execution · artifacts · reviews · control
node-specific context · record retrieval · shared tool documentation

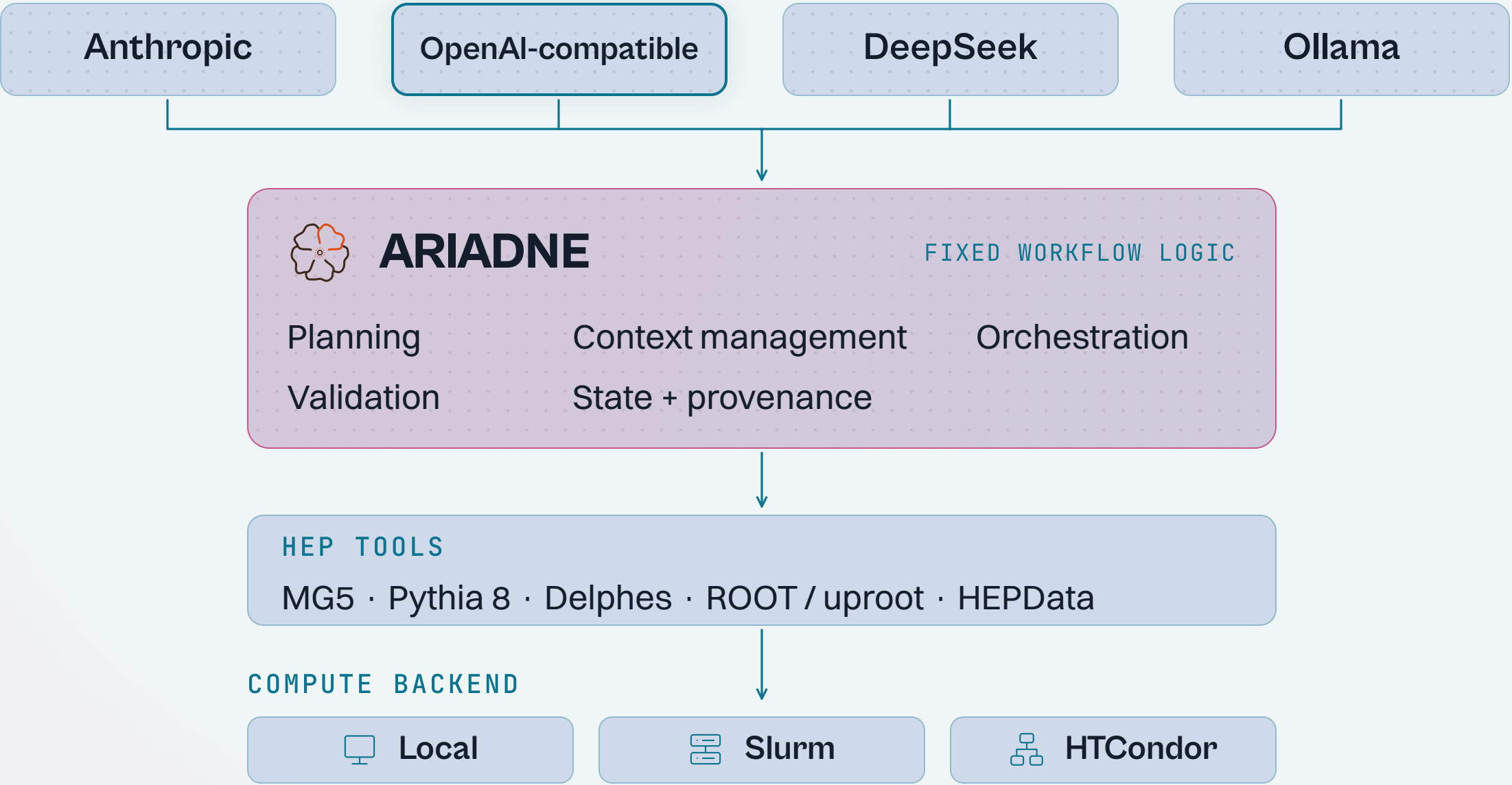
HEP domain knowledge



Domain-specific guidance helps the agent **follow HEP conventions**.

Model-independent **by design**

LLM BACKENDS



Models can be assigned by role

- Planner → Model A
- Coder → Model B
- Reviewer → Model A
- Research summary → Model C

Why separate the harness from the model?

- Benchmark the harness separately
- Use self-hosted or commercial models
- Trade capability against cost

The LLM, compute backend, and tools are **swappable components**, allowing tests to focus on the harness.

Ariadne

What we aim to improve

Consistently, across the full workflow.

 Long-horizon coherence

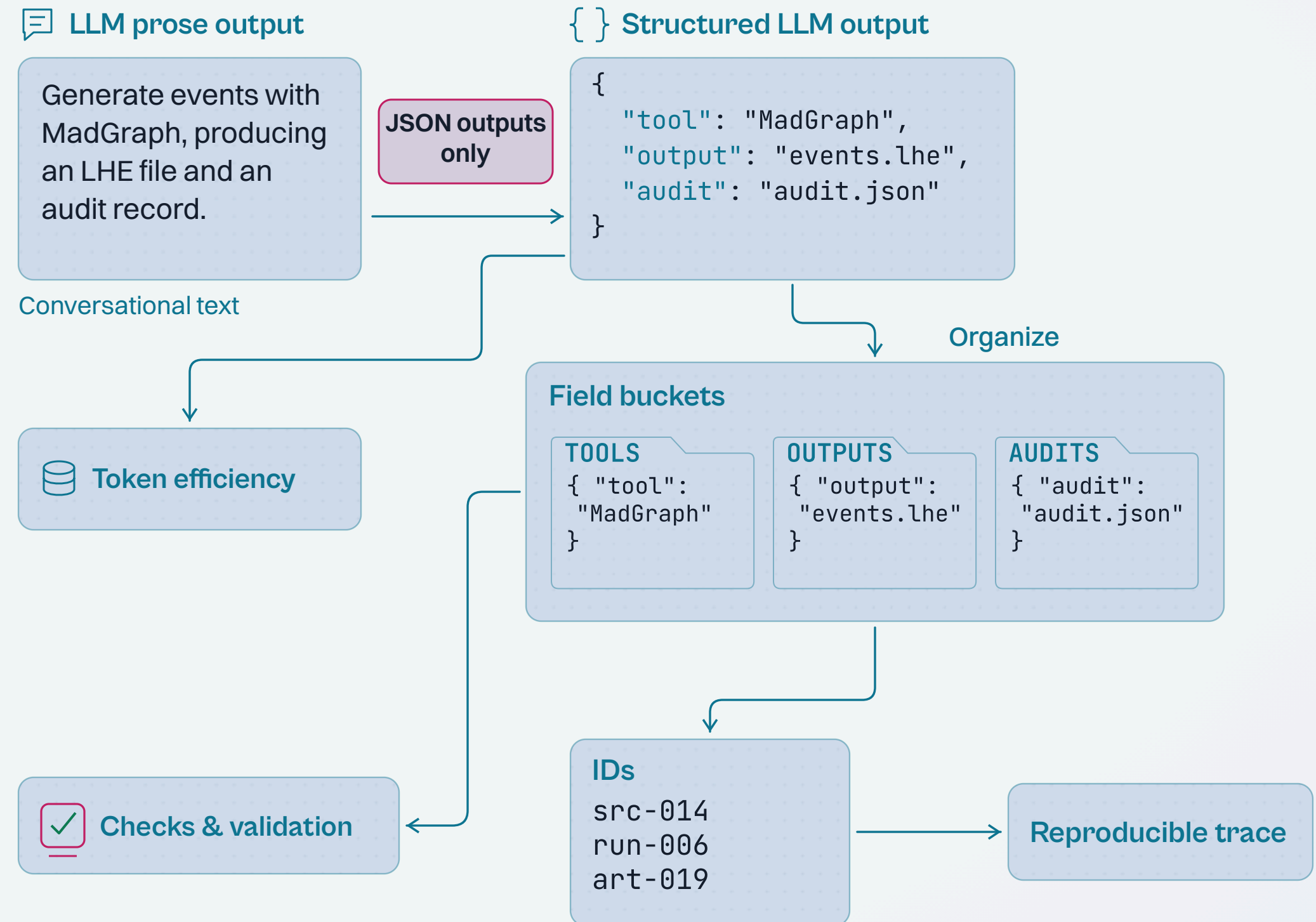
 Cost-efficient execution

 Reproducible and reliable

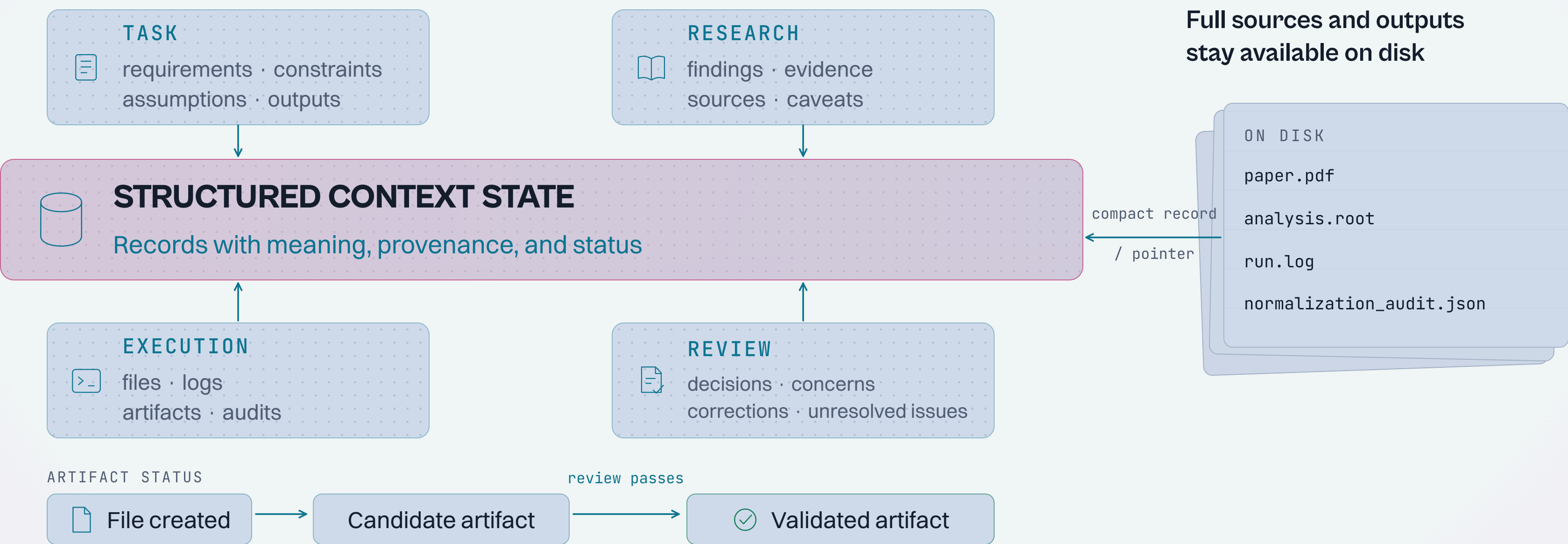
 LLM-agnostic

[1] Significant Gravitas (2023) · *Auto-GPT* · [GitHub](#) ↗

[2] OpenAI (2023) · *Function Calling and Other API Updates* · 13 June ↗

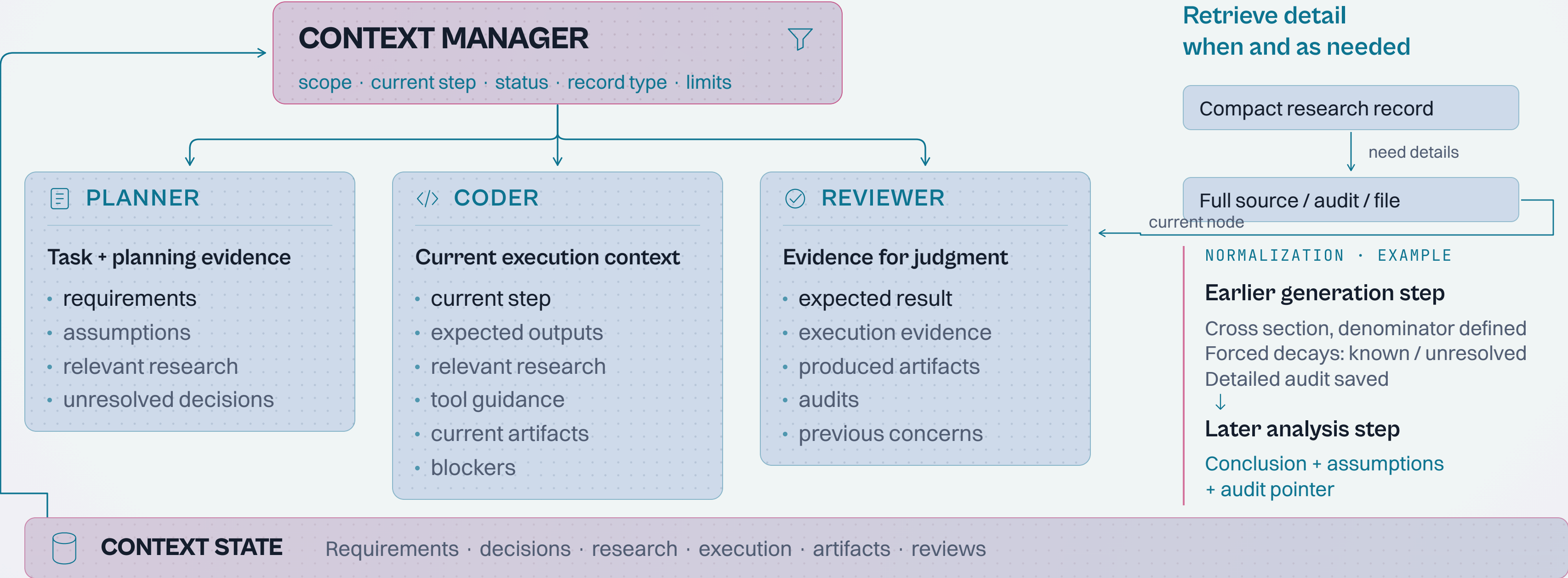


From conversation history to context state



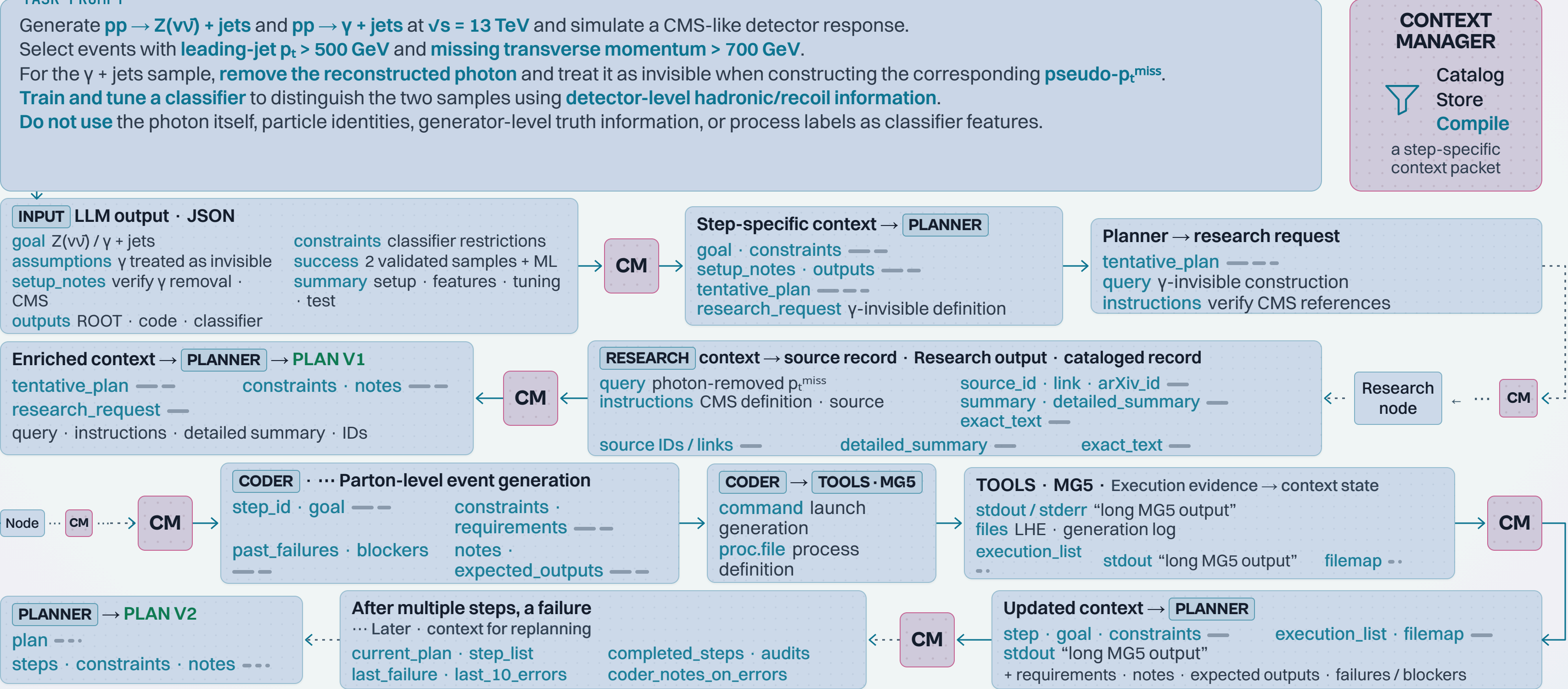
The workflow state stores **conclusions, evidence, and artifacts**—not just conversation history.

Curated context for each decision



The system remembers the workflow; **the model sees the relevant slice.**

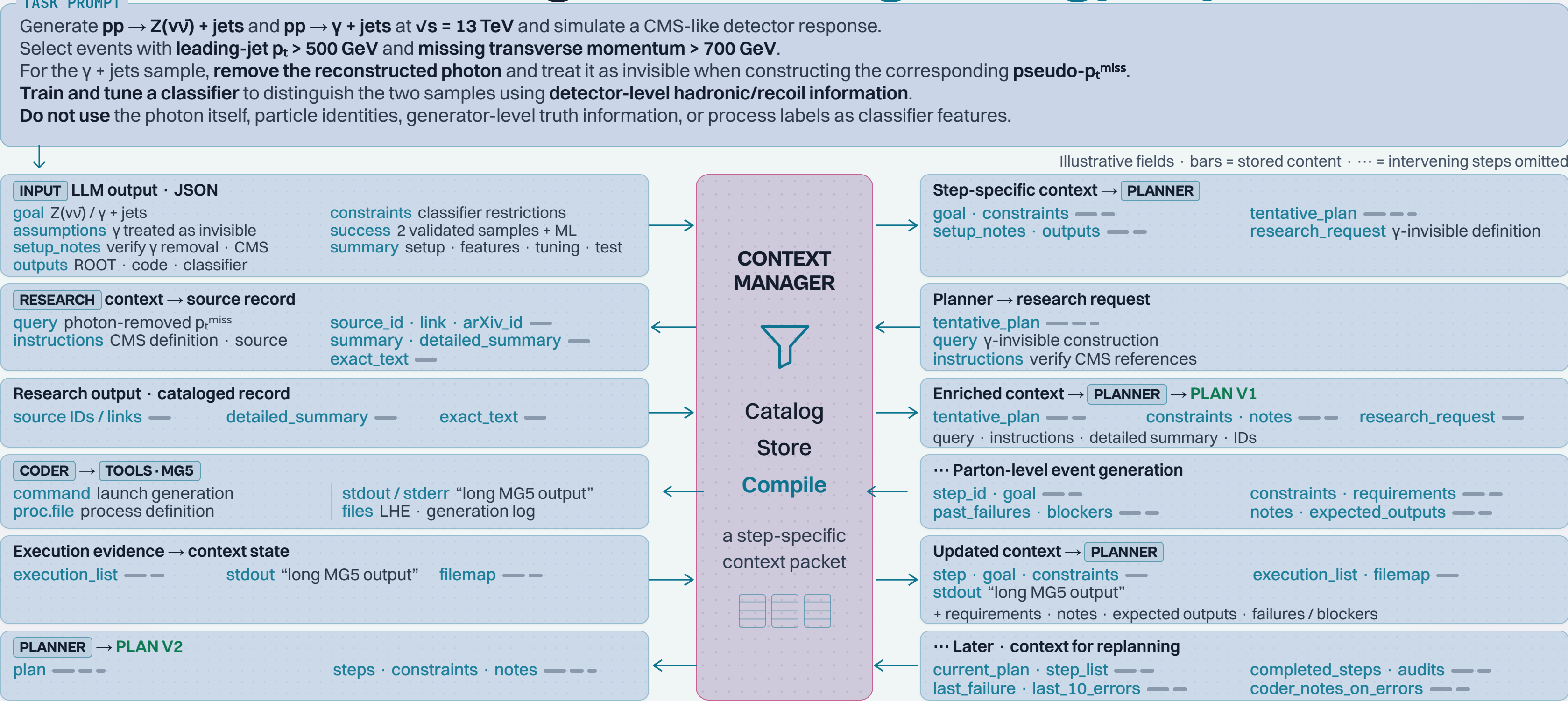
Custom Context Management for High Energy Physics



Illustrative fields · bars = stored content · dotted links = abbreviated workflow

Data is cleaned and cataloged; **the LLM sees a unique, relevant slice.**

Custom Context Management for High Energy Physics



Data is cleaned and cataloged; the LLM sees a unique, relevant slice.