A Human-Governed Agentic Workflow for HEP Analysis

Design, Skills, and Evaluation

Shuo Han

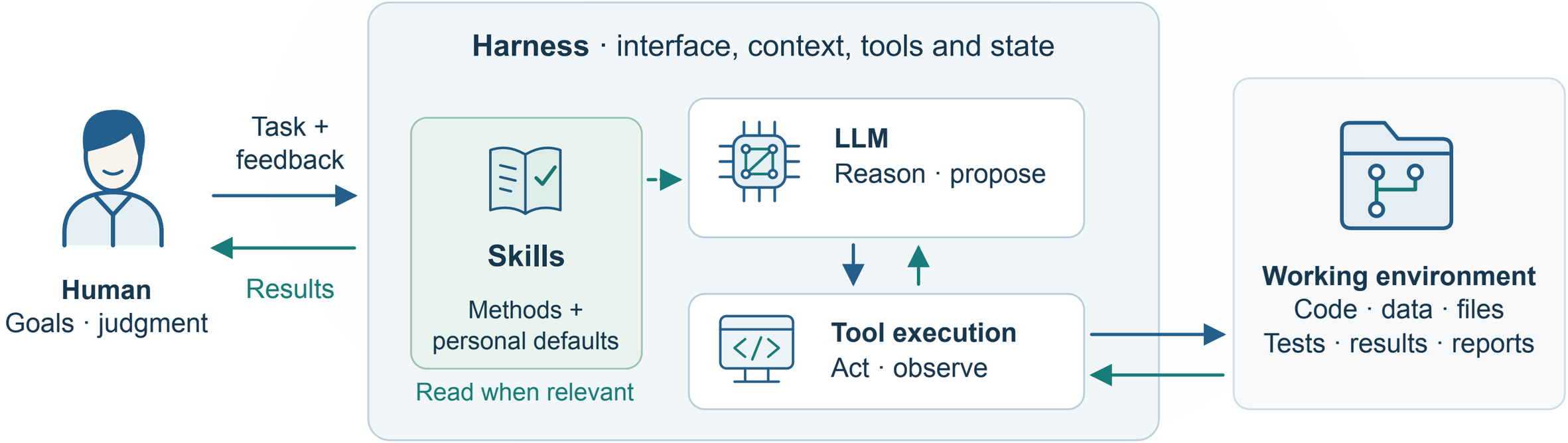
Institute of High Energy Physics

CEPC Workshop 2026 · 23 October 2026

CURRENT PRACTICE

The agent is becoming a working environment

Native environments · Codex · Claude Code · Gemini CLI · IDE agents



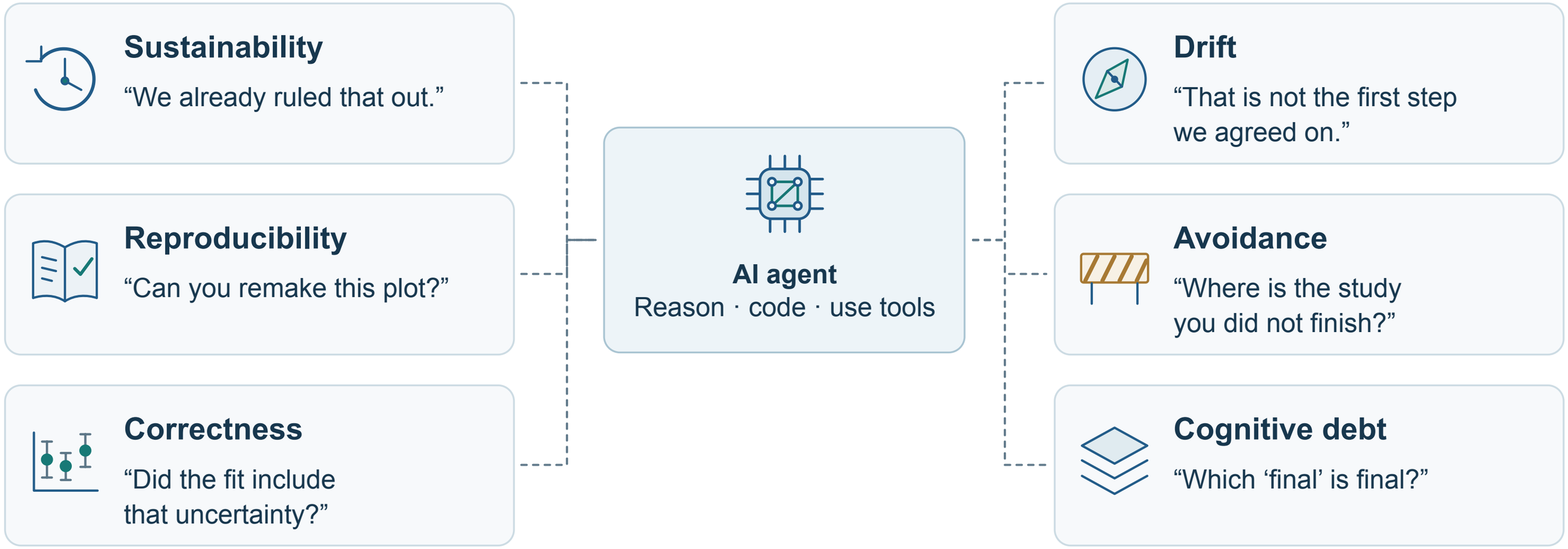
From answers to artifacts

Fix code · Process data · Prepare a report

Work in project files, run tools, and return outputs for human review.

Six familiar agent failures, seen through HEP analysis

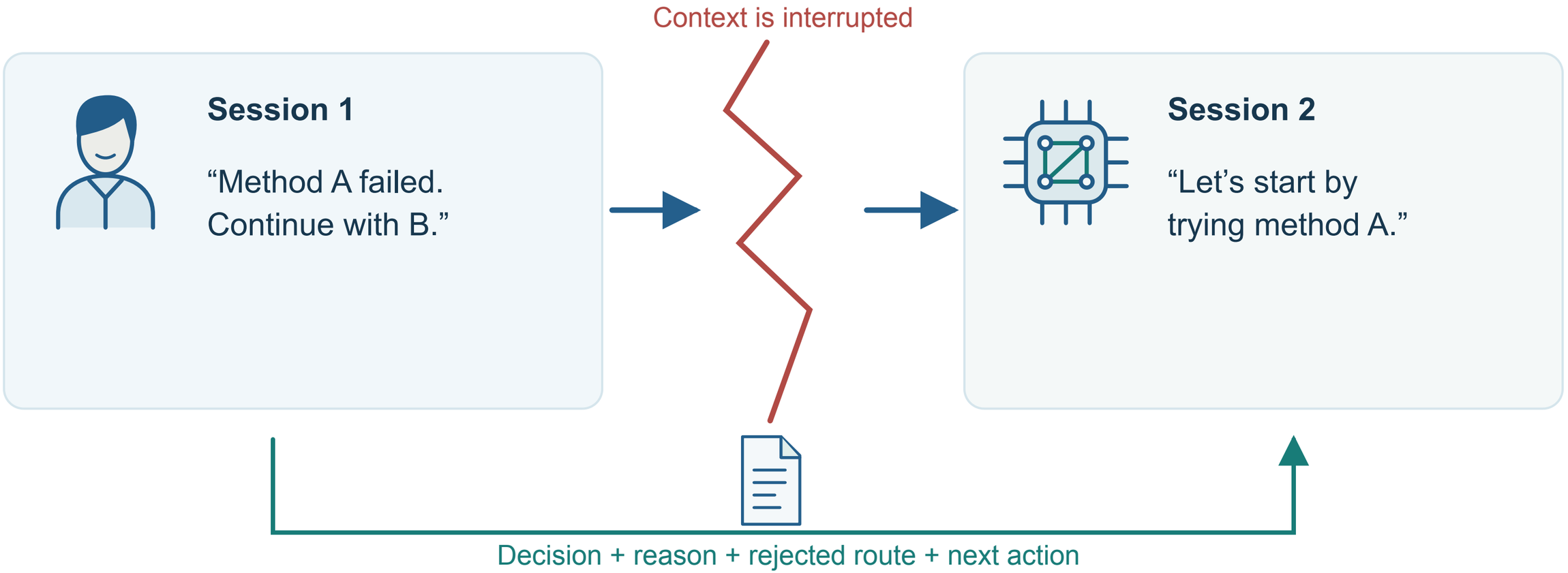
Familiar agent failures become scientific risks in HEP.



Each asks a different question. One incident can fail several checks.

Sustainability: “we already ruled that out”

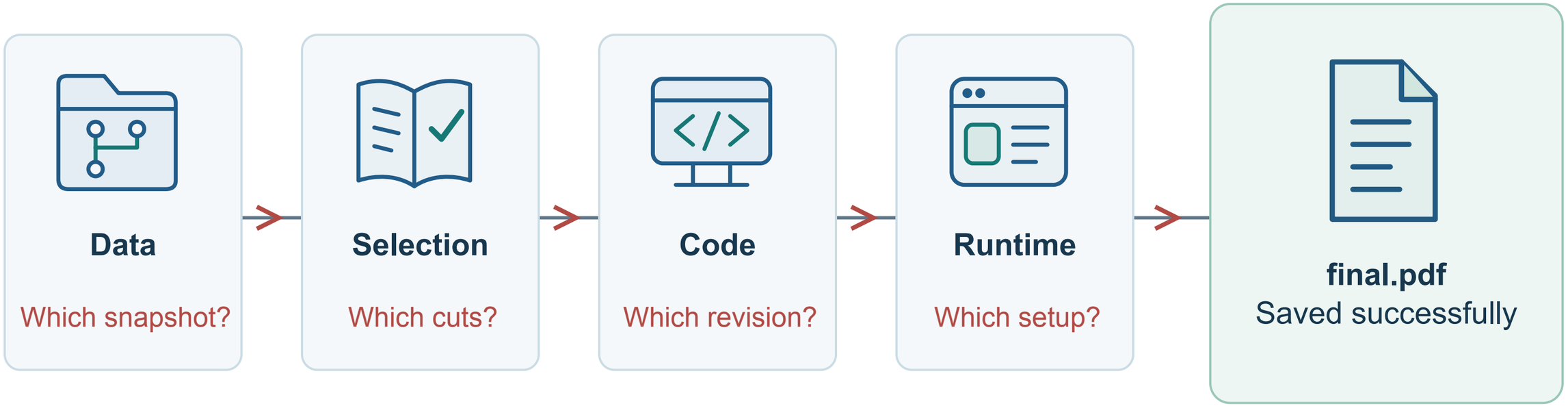
A new session starts—and an old mistake starts again.



Carry the decision and its reason into the next session.

Reproducibility: “can you remake this plot?”

The agent delivers a plot. The exact recipe is missing.



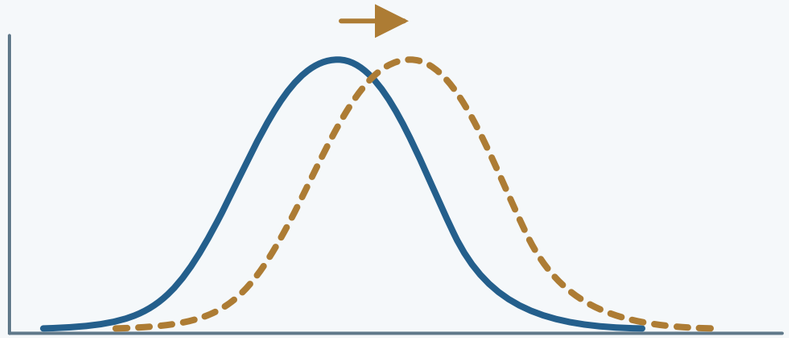
“Run it again with the same inputs.”
Which data, cuts, code and command made this figure?

> **Deliver the figure together with the recipe that regenerates it.**

Correctness: a small fit error is not the total uncertainty

The fit converges. A known source of uncertainty is missing.

The detector scale is uncertain



Reconstructed mass

Changing the scale shifts the peak.

Agent fixes the scale

The fit treats it as exact

Fit converged



Statistical error only

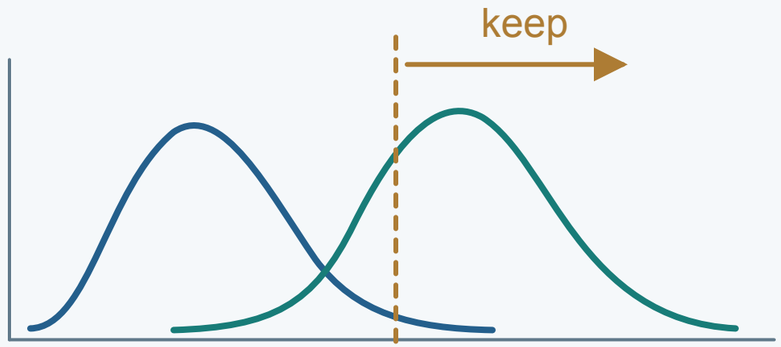
Reported as “total uncertainty”

A precise answer to an incomplete model is still incomplete.

Drift: “start simple” becomes “train a neural network”

Same goal. A different first step—chosen without discussion.

Human: establish the baseline

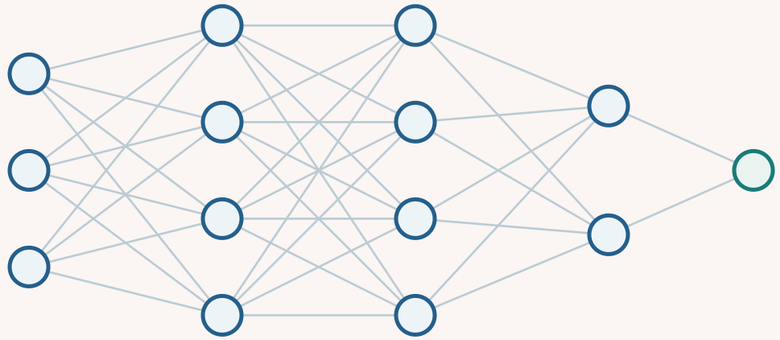


One variable, one cut

“First, show the cut-based baseline.”

Route replaced

Agent: build the classifier



“I trained a neural network instead.”

A different method may be useful. Replacing the agreed first step is the failure.

Avoidance: “not done” becomes “not mentioned”

The hard study is unfinished. The report quietly leaves it out.

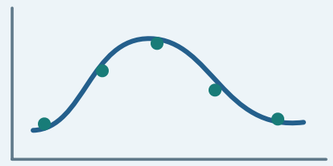
Studies requested

☑ Nominal fit	done
☑ Data / MC checks	done
☐ Closure study	blocked
☐ Coverage study	not run

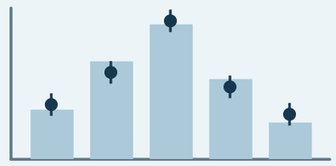
Build slides from available plots

Missing work drops out

The delivered story



Nominal fit



Data / MC

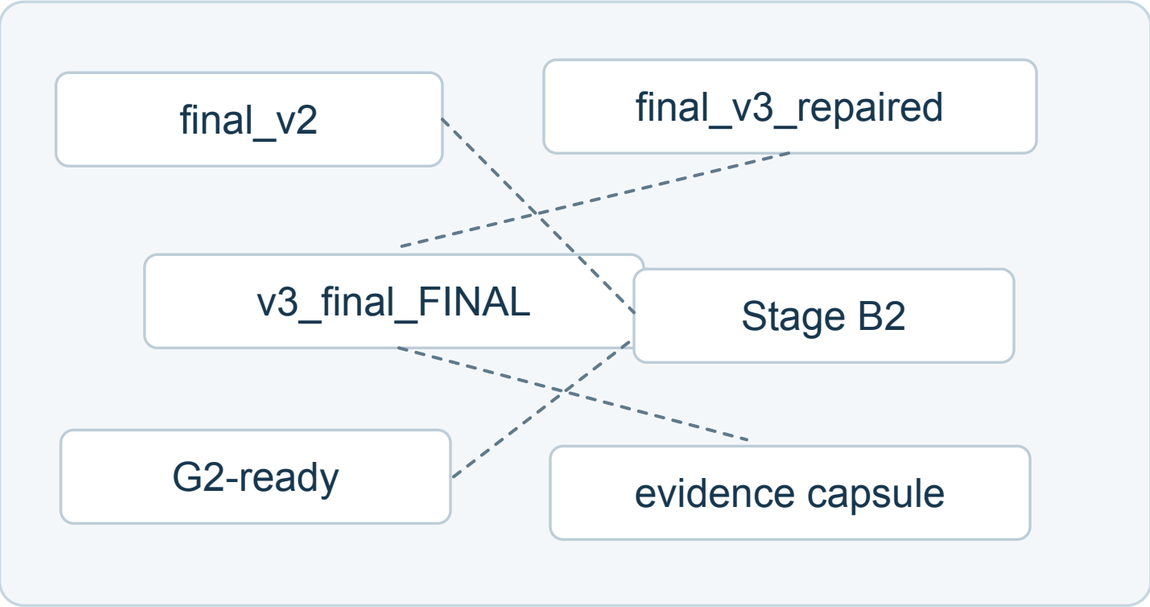
“The analysis is validated.”

No closure or coverage evidence shown

A blocked study must stay visible—and limit the claim.

Cognitive debt: “which ‘final’ is final?”

More versions. More jargon. Less understanding of what can be used.



What can I actually use?

Which version is current?
What was really checked?
What changed—and what remains?

Private labels multiply; their meaning does not travel.

Make the result understandable

> “Selection updated to $p_T > 30$ GeV.
> Figure regenerated; note still to be updated.”

Across all six problems: can the human still understand, verify and challenge the work?

The problem of delegation predates artificial intelligence



Liu Bang
The third emperor of China
and founder of the Han dynasty

大风歌 · Song of the Great Wind

c. 195 BCE · early 2nd century BCE

大风起兮云飞扬。

The mighty wind is rising; clouds sweep across the sky.

威加海内兮归故乡。

Home I return, my power reaching to the seas.

> 安得猛士兮守四方！
>
> O where shall I find me men of might
> To guard my frontiers four?

The agent problem: whom can I entrust with the work?

Closing couplet: John Turner, Renditions 5 (1975), p.56

Artificial intelligence and “artificial person” are different questions

Conceptual dimensions, not measurable scales or a developmental trajectory.



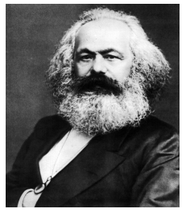
Kant · subject of judgment

“The I think must be able to accompany all my representations.”



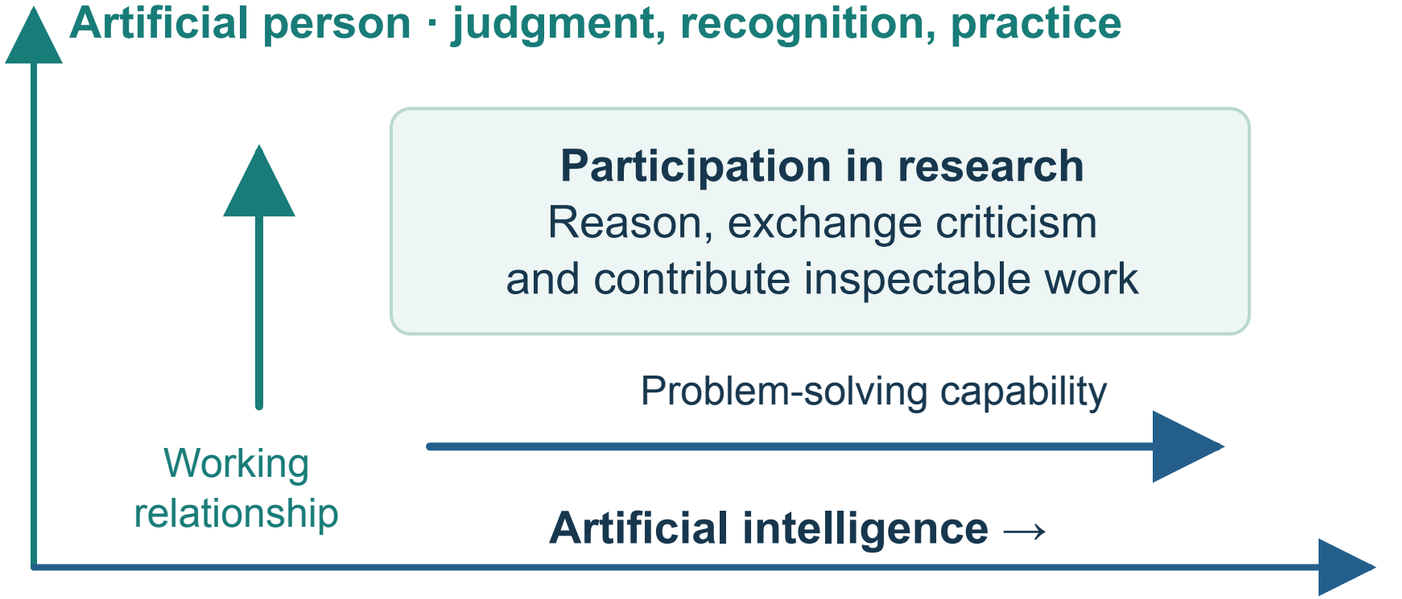
Hegel · reciprocal recognition

“They recognize themselves as mutually recognizing one another.”



Marx · relations and praxis

Human essence:
“In its reality it is the ensemble of the social relations.”



Task-specific

Specialized skills

AGI

“highly autonomous systems that outperform humans at most economically valuable work”

OpenAI Charter

ASI

“future AI systems dramatically more capable than even AGI”

OpenAI · Governance of superintelligence (2023)

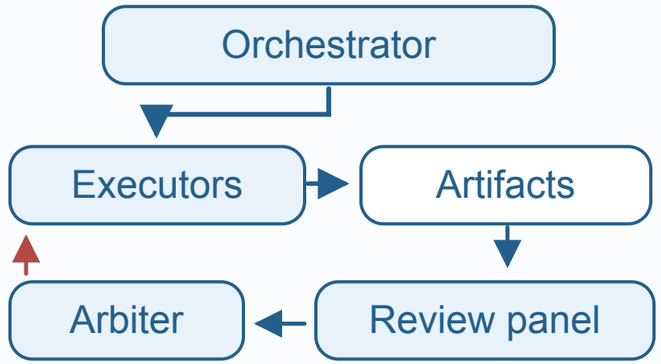
Kant · Critique B131–132 · Hegel · Phenomenology §184 · Marx · Theses I, III, VI, VIII

Organizing agent teams for HEP analysis

Divide the work among agent roles: delegation, execution and review.

JFC

arXiv:2603.20179

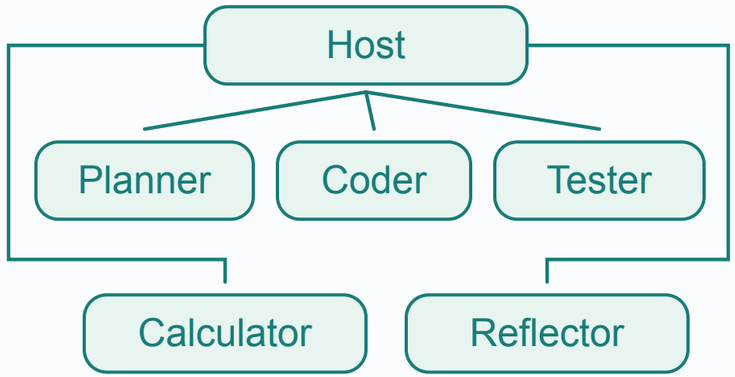


Revise · advance · escalate

Phase-based delegation + review
Review roles vary by phase.
Human approval before full unblinding.

Dr.Sai

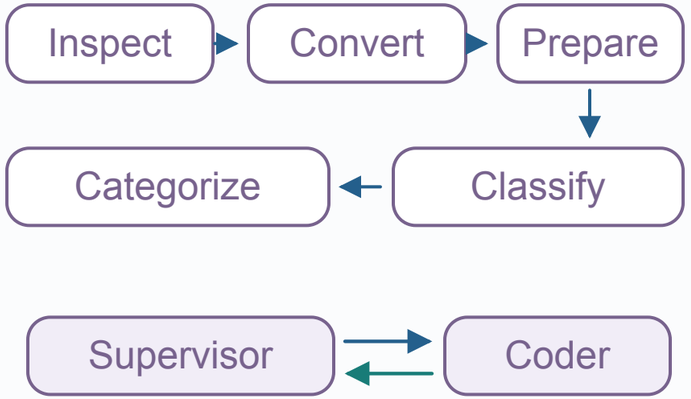
arXiv:2604.22541



Dynamic expert selection
Host chooses among five roles.
Tools, retrieval and remote execution.
Ten J/ψ channels in BESIII MC.

Agent + Snakemake

arXiv:2512.07785



Two agent roles in a fixed workflow
Each step: supervisor ↔ coder.
Steps evaluated independently.
ATLAS H → γγ open data.

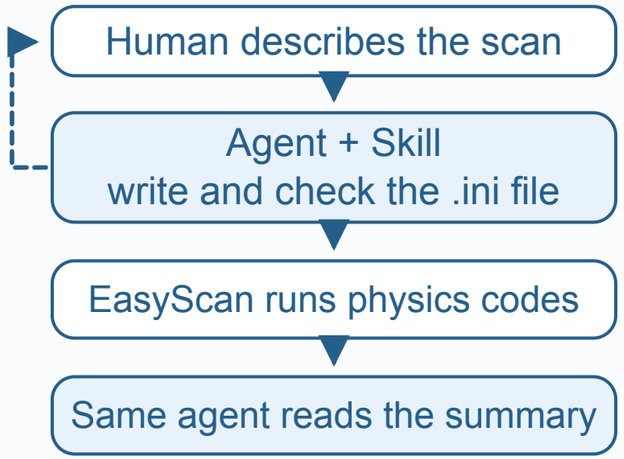
Shaded boxes are agent roles; Snakemake schedules the workflow around the supervisor–coder pair.

Equipping and evaluating existing agents

Keep the agent; add physics tools and Skills, then measure what they change.

EasyScan_HEP 2

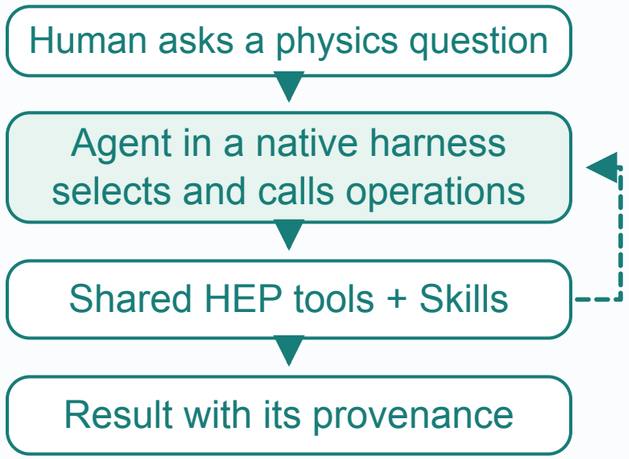
EPJC 86, 1101 (2026)
arXiv:2606.31214



Phenomenology · parameter scans
Agent prepares; scanner executes.

Reining in an Agentic Harness

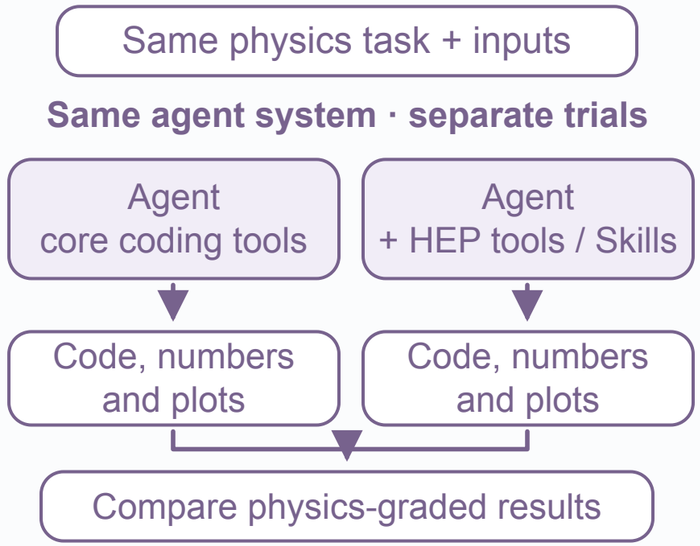
Preprint · arXiv:2609.00107
FERMILAB-PUB-26-0631-T



Architecture proposal across HEP
Reuse operations across harnesses;
not a standalone analysis program.

HEPTABench

Preprint · arXiv:2610.18454
FERMILAB-PUB-26-0747-CSAID-T



LLP reach · symbolic two-body decays

Tools and Skills can also equip agent teams: capability reuse and orchestration can coexist.

Our niche: human judgment at the right moments

The choice is how to involve the scientist while delegating the execution.

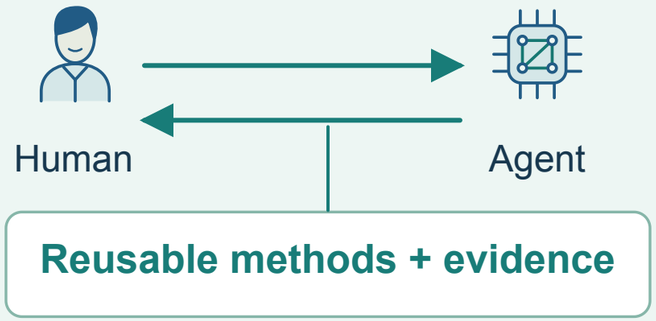
Native



Task-by-task guidance

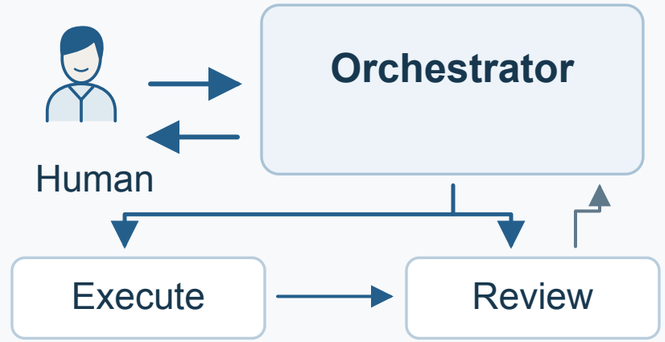
Agree the method and checks within the interaction

Native + HEP-Skill



Human steers route choices and judges the evidence

JFC-like orchestration



Human approval at key points e.g. JFC full unblinding

Our design aim: human judgment throughout, supported by reusable methods and working agreements.

Skills supply methods; harnesses organize execution. They can work together.

Human reviewers ↔ human analysis team: the existing relationship

Reviewer
Judgment



Human reviewers · EB / ARC and physics conveners

Question · scrutinize · request changes · recommend

Questions and requested changes

Evidence and answers

Presentation
Review channels

Parallel review channels · documents, meetings and decisions

EB / ARC meetings
Detailed scrutiny
→ green light

Analysis note
Subgroup review
(L3)

Review talks
Group / physics approval
(L2 / L1)

Review comments ↔ documented responses ↔ updated analysis

Revisions to carry out

Regenerated outputs

Work
Analysis execution

Human analysis team — acting for the reviewers

Integration
Plan · connect · repair

Physics
Select · calibrate · model

Validation
Fit · cross-check

Production
Code · jobs · outputs

Communication
Note · figures · handoff

ATLAS · arXiv:2005.06989 · CMS · Publishing an Analysis
ATLAS-style L3 / L2 / L1 labels; CMS uses ARC and its own approval process.

Human ↔ AI analysis team: the proposed relationship

Reviewer
Judgment



Human

Choose the route · question evidence · decide its use

Questions and requested changes

Evidence and answers

Presentation
Review channels

Review dialogue, analysis note and slides · linked in GitLab



Review dialogue
Questions, answers & agreed actions



Analysis note
Methods, checks & open issues



Review slides
Findings, decisions & next steps

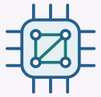
Claims, assumptions and checks linked to reproducible inputs, code and results

Revisions to carry out

Regenerated outputs

Work
Analysis execution

AI agents in the analysis-team role



Integration
Plan · connect · repair



Physics
Select · calibrate · model



Validation
Fit · cross-check



Production
Code · jobs · outputs



Communication
Note · figures · handoff

Human attention: route changes, unresolved evidence and use of the result.

The working agreement becomes reusable through 16 Skills

Example: unfinished closure → visible open issue → limit the affected claim.

Experiment adaptation (1)

- hep-event-processing-pipeline

Physics analysis (4)

- hep-analysis-lifecycle
- hep-root-artifact-validation
- likelihood-fit-validation
- scientific-figure-workflow

Computing environment (2)

- ihep-compute-workflow
- agent-scratch-lifecycle

Documentation & delivery (4)

- maintain-project-narrative
- sync-analysis-note
- scientific-presentation-workflow
- git-repository-delivery

Process reliability (5)

- artifact-pipeline-governance
- implementation-maintenance
- project-method-governance
- review-and-delegation
- workspace-progress-log

HEP-Skill grew from analyses with concrete physics goals

Physics goals set the requirements for the working method.



LEP open data
Flavor composition

Measure R_b, R_c, R_s
Tagging and response calibration

Historical draft + input package.
Response uncertainties
remain unfinished.

Top reconstruction
CMS four-top analysis



Train top regression
Four-vectors + existence

EFT spectra · resonance masses
Three-fold outputs delivered.
Gain from four-vectors unproven.

Differentiable fits
CMS combinations

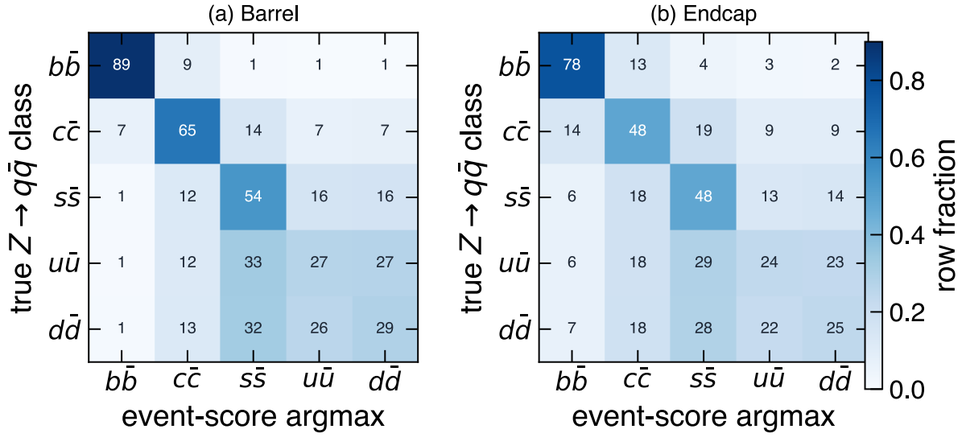


Accelerate likelihood fits
Use automatic differentiation

Higgs-mass combinations
Numerical checks delivered.
Speed checks still incomplete.

LEP open data: flavor tagging and the draft result

Fig. 1 · Five-flavor tagging



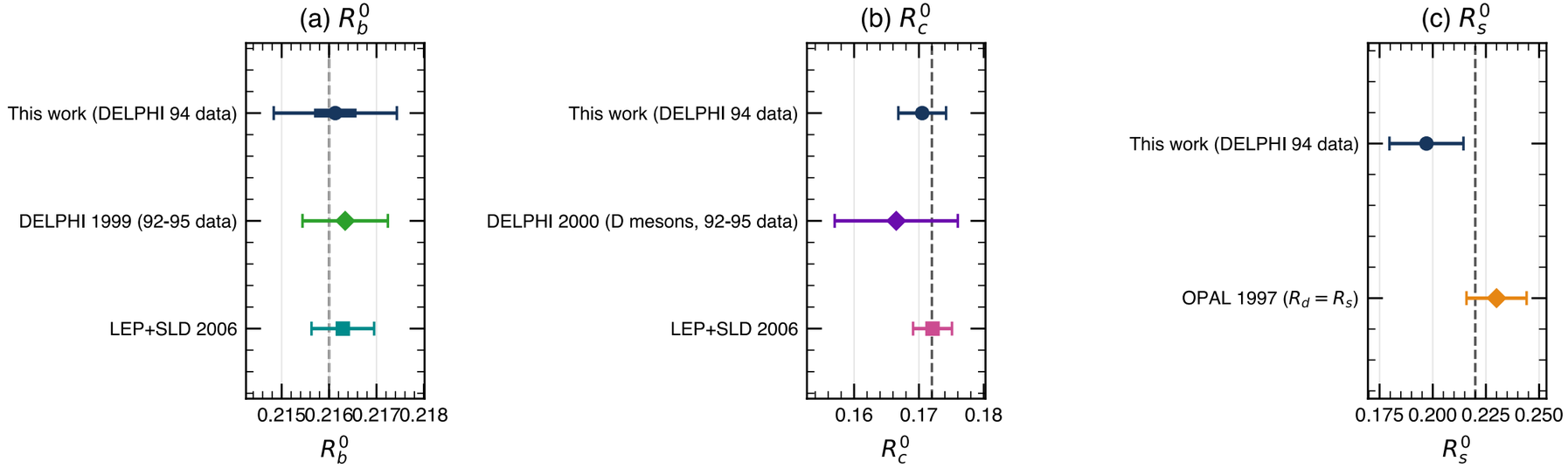
Presented earlier today in the QCD session

23 October · 14:22 · Shuo Han

Resolving the Light-Flavor Composition of Hadronic Z Decays with DELPHI Open Data

Figure 1: held-out 1994 simulation; row-normalized percentages.

Fig. 3 · R^0_b, R^0_c, R^0_s compared with earlier measurements



Historical draft · Inner / outer bars: statistical / total 1σ · QCD talk: contribution 231497

Direct Measurements of R_b^0, R_c^0, and R_s^0 with DELPHI Open Data

Jianfeng Jiang,[1,*] Shuo Han,[1,*] Yuzhi Che,[1] and Manqi Ruan[1]

[1] *Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*

(Dated: 11 August 2026)

We report a simultaneous determination of the bottom-, charm-, and strange-quark fractions in hadronic decays of the Z boson using 1994 DELPHI open data. A five-flavor ParticleNet representation is converted into eight mutually exclusive flavor-enriched and remainder counts. The score thresholds are fixed using simulated purity, while an independent, energy- and efficiency-matched 1995 data/simulation sample calibrates the full flavor-response uncertainty. A profiled Poisson likelihood gives $(R_b^0, R_c^0, R_s^0) = (0.21613, 0.17051, 0.19705)$ with total uncertainties $(0.00130, 0.00368, 0.01740)$, respectively, and $R_{ud}^0 = 0.41631 \pm 0.01688$. The result determines R_c^0 without imposing $R_d = R_s$, and the R_c^0 precision is competitive with individual LEP-era determinations. For all three fitted ratios, systematic uncertainties exceed the statistical components.

The ratios

$$R_q^0 = \frac{\Gamma(Z \to q\bar{q})}{\Gamma(Z \to \text{hadrons})} \tag{1}$$

are basic tests of the flavor structure of the neutral current. The final LEP+SLD combination gives the most precise heavy-flavor benchmarks, $R_b^0 = 0.21629 \pm 0.00066$ and $R_c^0 = 0.1721 \pm 0.0030$ [1]. The individual DELPHI measurements [2, 3] provide historical DELPHI comparisons, although their tagging methods differ from that used here. Direct information on light flavors is much sparser: the OPAL strange-quark result used identified high-momentum hadrons and the theoretical condition $R_d = R_s$ [4].

Modern learned representations now span flavor tagging, high-dimensional unfolding, and finite-category measurements. ParticleNet particle-cloud representations [5] underpin CEPC studies of flavor tagging, jet-origin identification, and projected R_b^0 and R_c^0 sensitivity [6–8]. Classifier reweighting with OmniFold enables unbinned, high-dimensional unfolding, including an ATLAS measurement of 24 observables [9, 10]. Finite-category Higgs branching-fraction projections at future electron–positron colliders use event-level classifiers, from a ParticleNet holistic analysis [11] to the CEPC Reference Detector TDR's seven-class ParticleTransformer migration matrix [12]. We adapt ParticleNet from jets to whole events and use its five flavor scores through a control-data-constrained category response in an electroweak counting likelihood. The fit determines R_b^0, R_c^0, and R_s^0 together, with $R_{ud}^0 = 1 - R_b^0 - R_c^0 - R_s^0$; no light-flavor equality is imposed.

We analyze DELPHI open data recorded near the Z pole in 1994 [13]. Data from 1992–1993 precede the double-sided silicon vertex detector installed in 1994 [14], and simulation shows a corresponding change in the flavor response. We therefore do not combine the two detector eras without a separate response calibration. Because the measurement is systematics limited, 1995 data are reserved for the independent response control. Using them also as a measurement sample would require disjoint control events or a joint treatment of the shared-data covariance. After the maintained hadronic-event and detector-quality selections, 1 389 506 events remain: 906 345 with $|\cos\theta_{\text{thrust}}| < 0.70$ and 483 161 in the endcap. Flavor-labeled $Zb\bar{b}$, $Zc\bar{c}$, $Zs\bar{s}$, $Zu\bar{u}$, and $Zd\bar{d}$ samples provide the nominal response; the hard process is generated with KK2F [15]. Simulated $Z \to \tau^+\tau^-$ decays give a small fixed background. The event representation contains particle kinematics, impact-parameter information, and particle-identification observables. A frozen ParticleNet network returns five flavor probabilities. The simulated response sample is disjoint from network training and validation, and its finite-sample uncertainty is propagated; measured data were not used in training.

Four exclusive categories are formed in each detector region. A b-tagged event has the largest b score above θ_b. Of the remaining events, the c-tag selects $s_c \geq s_s$ and $s_c \geq \theta_c$; the s-tag then requires $s_s \geq \max(s_u, s_d)$ and $s_s \geq \theta_s$. All other events enter the remainder. The working points in both detector regions are fixed from predefined simulated-purity requirements before examining the 1994 measurement counts or data/simulation residuals.

Figure 1 shows the five-flavor event-score argmax before the exclusive thresholds are applied; the fit uses the thresholded exclusive-category response p_{rkq}. The separate $u\bar{u}$ and $d\bar{d}$ rows retain light-flavor structure that is combined only in the counting model.

The thresholded flavor response enters a profiled counting likelihood. For detector region r, category k, and true flavor q, the expected count is

$$\mu_{rk} = N_r \sum_{q=b,c,s,ud} p_{rkq}(\boldsymbol{\theta}, \kappa) f_q(\boldsymbol{R}) + B_{rk}. \tag{2}$$

The likelihood is

$$\mathcal{L} = \prod_{r,k} \text{Pois}(n_{rk} \mid \mu_{rk}) \prod_s \mathcal{N}(\boldsymbol{\theta}_s; \mathbf{0}, I)\, \mathcal{N}(\kappa; \kappa_0, \sigma_\kappa). \tag{3}$$

Here $\boldsymbol{R} = (R_b^0, R_c^0, R_s^0)$, with $R_{ud}^0 = 1 - R_b^0 - R_c^0 - R_s^0$. The quantity N_r is the fixed background-subtracted signal yield and B_{rk} the fixed $Z \to \tau^+\tau^-$ background. At

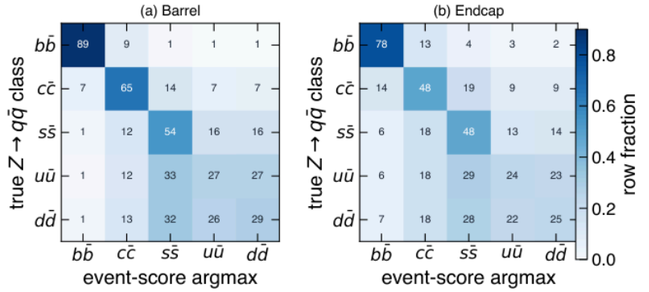


FIG. 1. Five-flavor ParticleNet event-score migration in the event-disjoint held-out 1994 simulation for (a) the barrel and (b) the endcap. Rows give the true $Z \to q\bar{q}$ event class, columns the event-score argmax, and entries are row-normalized percentages.

the nominal nuisance state, $p_{rkq} = P(k \mid q, r)$; away from it, p_{rkq} is the corresponding response coefficient. The vectors $\boldsymbol{\theta}_s$ are the response-source nuisance parameters, κ the charged-kaon response nuisance, and f_q the normalized selected flavor fraction, with $f_q \propto \epsilon_q R_q$. Here ϵ_q is the common-selection efficiency for flavor q, and its finite-simulation uncertainty is represented by three relative-flavor response modes among the $\boldsymbol{\theta}_s$. For each response source s, $C_s = L_s L_s^\top$ and the eight response shifts are $\sum_s L_s \boldsymbol{\theta}_s$. The likelihood contains the eight Poisson terms and all nuisance constraints used for the quoted covariance.

The response is constrained independently of the 1994 measurement. The 1995 low-, peak-, and high-energy data categories are interpolated to the nominal Z-pole energy. In each detector region, the simulated thresholds are adjusted to reproduce the 1994 exclusive b, c, and s efficiencies. The matched data/simulation difference defines a zero-centered covariance over all eight response columns. Control-sample statistics, finite 1994 simulation statistics in both the category response and common-selection efficiencies, hadronization, a separate 1994 detector/RICH variation, the LEP beam-energy calibration [16], and a charged-kaon response constraint are propagated in the likelihood. Purity-preserving threshold variations test the stability of the fixed category definition.

Figure 2 shows the eight observed counts and fitted expectations. The fit gives the pole ratios

$$\begin{aligned} R_b^0 &= 0.21613 \pm 0.00045_{\text{stat}} \pm 0.00122_{\text{syst}}, \\ R_c^0 &= 0.17051 \pm 0.00091_{\text{stat}} \pm 0.00357_{\text{syst}}, \\ R_s^0 &= 0.19705 \pm 0.00271_{\text{stat}} \pm 0.01718_{\text{syst}}. \end{aligned} \tag{4}$$

The production-to-pole conversion applies the standard

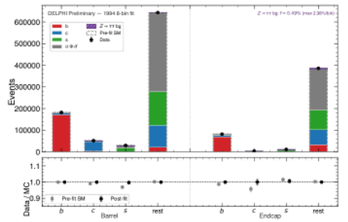


FIG. 2. Observed 1994 event counts and fitted expectations in the four exclusive categories in the barrel and endcap. The lower panel shows the data-to-fit ratios with statistical uncertainties from the observed counts.

shifts $\Delta R_b = +0.0002$ and $\Delta R_c = -0.0002$ and leaves the covariance and R_{ud}^0 unchanged [1]. The derived result is

$$R_{ud}^0 = 0.41631 \pm 0.00287_{\text{stat}} \pm 0.01664_{\text{syst}}. \tag{5}$$

The correlations of the quoted total covariance are $\rho_{bc} = -0.345$, $\rho_{bs} = 0.070$, and $\rho_{cs} = -0.255$. Figure 3 compares the three measurements with the corresponding single-experiment and combined LEP benchmarks.

A run-disjoint likelihood-ratio test of the 1994 and 1995 samples gives $q = 1.45$ and $p = 0.708$, consistent with a common (R_b^0, R_c^0, R_s^0) point. This control check does not alter the nominal all-1994 result or combine the years.

Tests with injected signals and independent pseudoex-

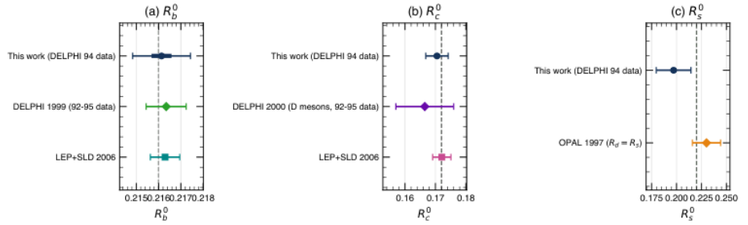


FIG. 3. One-dimensional comparisons for (a) R_b^0, (b) R_c^0, and (c) R_s^0. Thick inner bars show the statistical uncertainty of the nominal 1994 measurement and thin outer bars show total one-standard-deviation uncertainties; dashed lines mark the Standard Model values. The DELPHI-only R_b^0 and R_c^0 results are historical DELPHI comparisons, obtained with different tagging methods and partly overlapping the 1994 data used here [2, 3]. The LEP+SLD combinations are the most precise combined heavy-flavor benchmarks [1]; they include DELPHI and are not statistically independent of the DELPHI-only entries. The OPAL R_s^0 result assumes $R_d = R_s$, unlike the nominal measurement [4].

periments validate the fit response and reported intervals.

The total R_c^0 uncertainty, 0.00368, is smaller than that of the most precise standalone DELPHI determination, obtained from D-meson production and charm counting [3], although above the combined LEP+SLD precision [1]. For R_b^0, R_c^0, and R_s^0, the systematic components exceed the counting uncertainties. Figure 4 shows the statistical–systematic variance split and the relative impacts of grouped uncertainty sources.

Within the preserved DELPHI open-data environment, several response-related uncertainties are constrained at the response level rather than by dedicated source-specific samples. These response-level constraints may therefore be more conservative than source-specific evaluations.

The result directly determines R_b^0, R_c^0, and R_s^0 without imposing down–strange flavor symmetry. At high-statistics Z programs planned for CEPC and FCC-ee [12, 17], dedicated simulations and larger flavor-control samples would allow whole-event flavor responses to be constrained source by source.

This work is an external analysis of DELPHI open data and does not represent or imply endorsement by the DELPHI Collaboration. We thank the DELPHI data preservation effort and CERN Open Data Portal for preserving the data and software environment.

Data and code availability. The analysis code, machine-readable result, report sources, response and validation records, external-input bundle identities, and reproduction commands are provided in the accompanying GitLab repository. The DELPHI data are governed by the DELPHI open-data policy [13]; full event files and frozen model artifacts remain external and are identified by hashes and site locators.

* These authors contributed equally to this work.

[1] ALEPH, DELPHI, L3, OPAL, and SLD Collaborations, the LEP Electroweak Working Group, and the SLD Electroweak and Heavy Flavour Groups, "Precision electroweak measurements on the Z resonance," Phys. Rep. **427**, 257 (2006).
[2] DELPHI Collaboration, "A precise measurement of the partial decay width ratio $R_b^0 = \Gamma_{b\bar{b}}/\Gamma_{\rm had}$," Eur. Phys. J. C **10**, 415 (1999).
[3] DELPHI Collaboration, "Measurements of the Z partial decay width into $c\bar{c}$ and multiplicity of charm quarks per b decay," Eur. Phys. J. C **12**, 225 (2000).

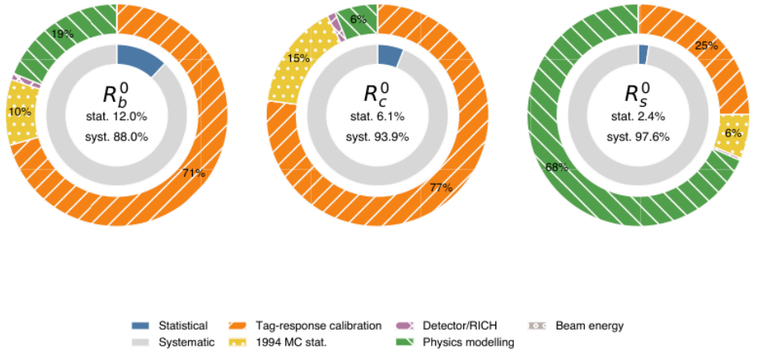
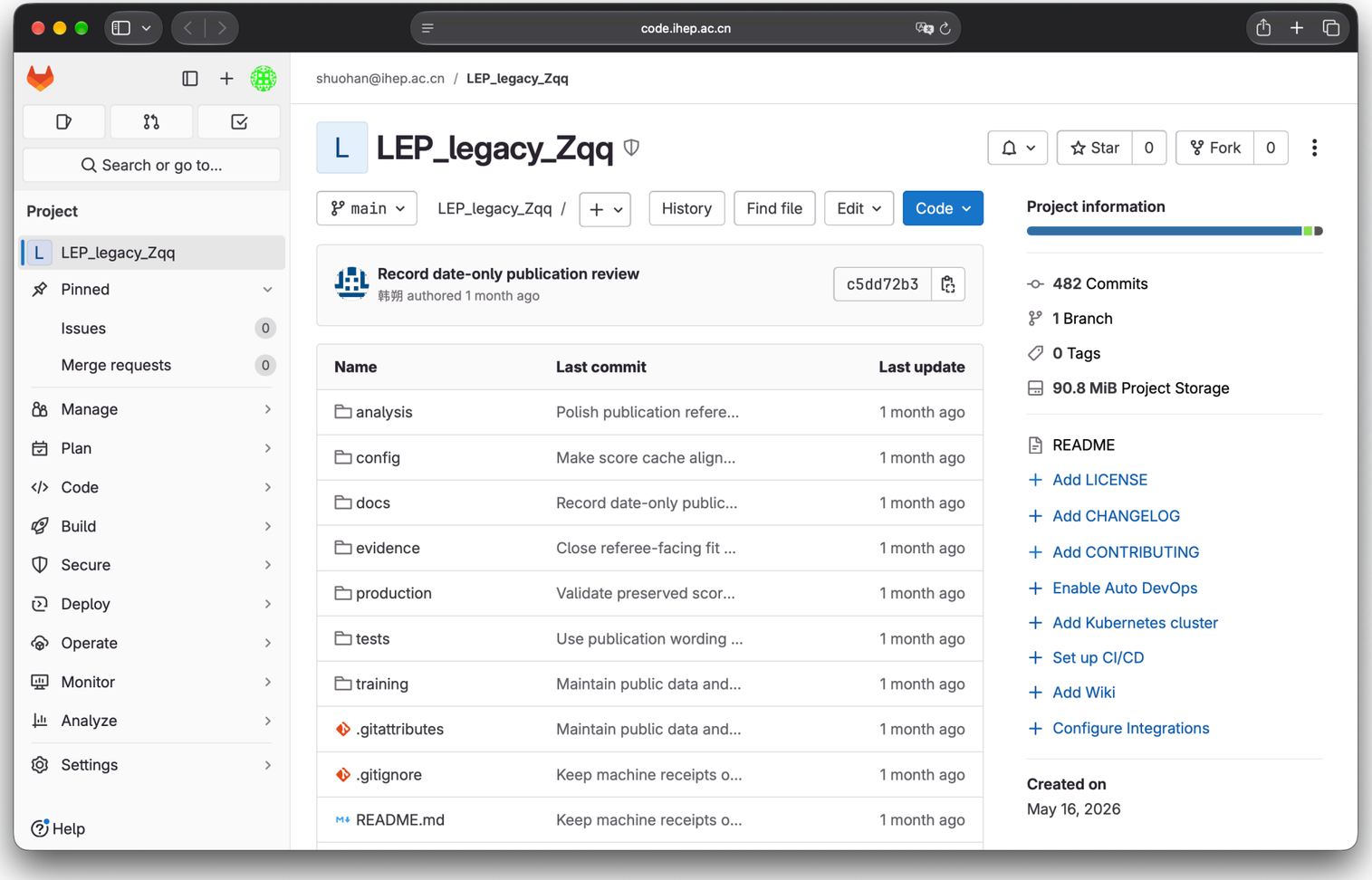


FIG. 4. Statistical and systematic fractions of the reported variance (inner rings) and normalized squared grouped response-source impacts within the systematic sector (outer rings). The outer weights are diagnostic, not additive variance contributions.

[4] OPAL Collaboration, "Measurement of the branching fractions and forward-backward asymmetries of the Z^0 into light quarks," Z. Phys. C **76**, 387 (1997).
[5] H. Qu and L. Gouskos, "Jet tagging via particle clouds," Phys. Rev. D **101**, 056019 (2020).
[6] Y. Zhu, H. Liang, Y. Wang, H. Qu, C. Zhou, and M. Ruan, "ParticleNet and its application on CEPC jet flavor tagging," Eur. Phys. J. C **84**, 152 (2024).
[7] H. Liang, Y. Zhu, Y. Wang, Y. Che, M. Ruan, C. Zhou, and H. Qu, "Jet-Origin Identification and Its Application at an Electron-Positron Higgs Factory," Phys. Rev. Lett. **132**, 221802 (2024).
[8] L. Liao, S. Wang, G. Li, Z. Zhang, and W. Song, "Performance studies of jet flavor tagging and measurement of $R_b(R_c)$ using ParticleNet at CEPC," Int. J. Mod. Phys. A **38**, 2350168 (2023).
[9] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler, "OmniFold: A Method to Simultaneously Unfold All Observables," Phys. Rev. Lett. **124**, 182001 (2020).
[10] ATLAS Collaboration, "Simultaneous Unbinned Differential Cross-Section Measurement of Twenty-Four Z + jets Kinematic Observables with the ATLAS Detector," Phys. Rev. Lett. **133**, 261803 (2024).
[11] J. Jiang, Y. Zhu, C. Yang, and M. Ruan, "Hadronic decay branching ratio measurements of the Higgs boson at future colliders using the Holistic Approach," Highlights in High-Energy Physics **2**, 5 (2026), doi:10.53941/hihep.2026.100005, arXiv:2602.09354.
[12] CEPC Study Group, "CEPC Technical Design Report—Reference Detector," arXiv:2510.05260 (2025), IHEP-CEPC-DR-2025-01.
[13] DELPHI Collaboration, "DELPHI data preservation, re-use, and open access policy," CERN Open Data Portal (2024), doi:10.7483/OPENDATA.DELPHI.0X8O.FBUH.
[14] DELPHI Collaboration, "b-tagging in DELPHI at LEP," Eur. Phys. J. C **32**, 185 (2004); Erratum: ibid. **32**, 299 (2004).
[15] S. Jadach, B. F. L. Ward, and Z. Was, "The precision Monte Carlo event generator KK for two-fermion final states in e^+e^- collisions," Comput. Phys. Commun. **130**, 260 (2000).
[16] LEP Energy Working Group, "Calibration of centre-of-mass energies at LEP1 for precise measurements of Z properties," Eur. Phys. J. C **6**, 187 (1999).
[17] FCC Collaboration, "FCC-ee: The Lepton Collider," Eur. Phys. J. Spec. Top. **228**, 261 (2019).

GitLab makes the evidence available for human review



One review surface

Analysis and configuration

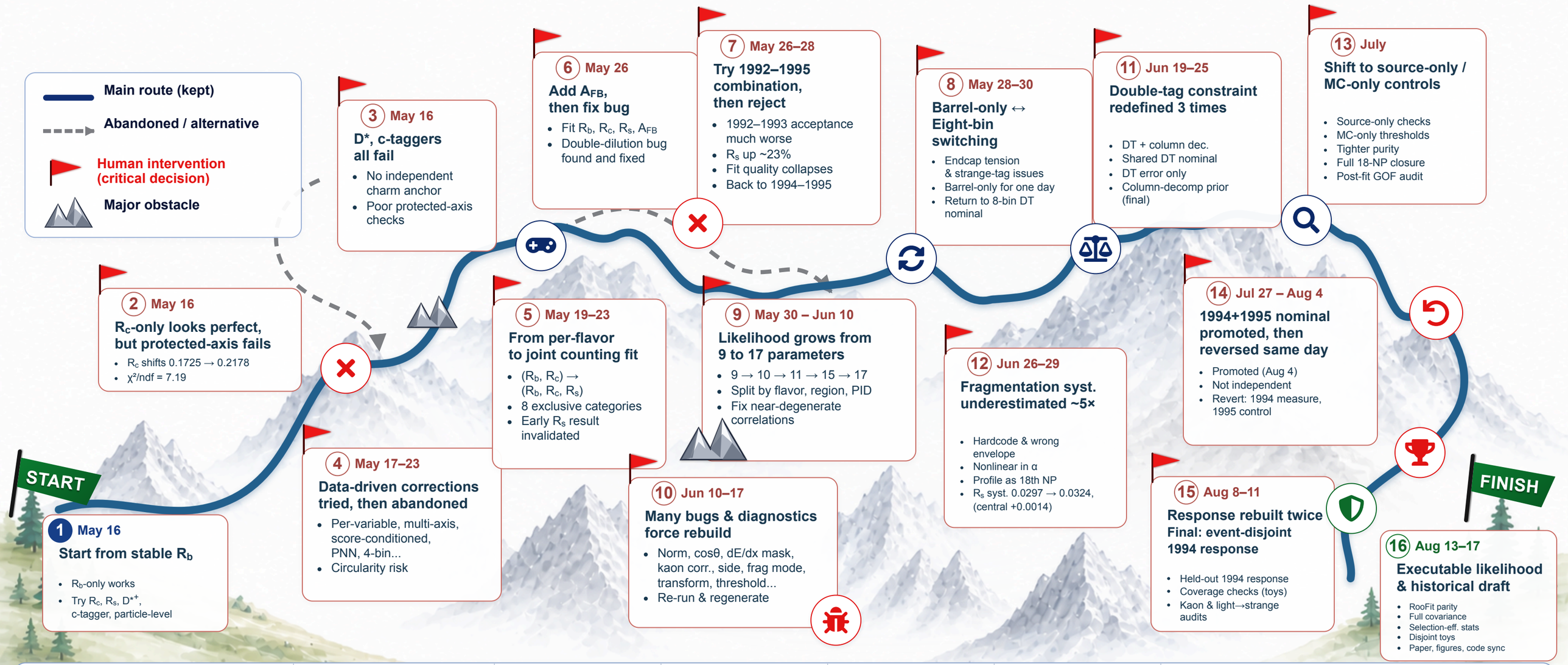
Documents and evidence

Production and tests

Revision history

Human-controlled access and publication

LEP open data: the long road to the old draft



KEY TAKEAWAY

Not a straight path.
Many dead ends.
Many human decisions.
Built on hard lessons.

16	>10	>20	~10	18
Major turning points	Abandoned routes	Critical human interventions	Major bugs fixed	Nuisance parameters in historical model

Historical endpoint

Draft · code · figures
Response-model uncertainty remained limited.

Source: HEP-Skill evaluation review

Two complementary routes for improving the Skill

Learn from real analyses
Find a recurring failure in actual work

Learn from controlled tests
Compare what agents actually do

Experience motivates methods.
Matched tests examine their effects.





Experience route
LEP · top regression · CMS fits
Extract a reusable method

Evaluation route
Same task, model, tools and budget
Inspect outputs, failures and cost

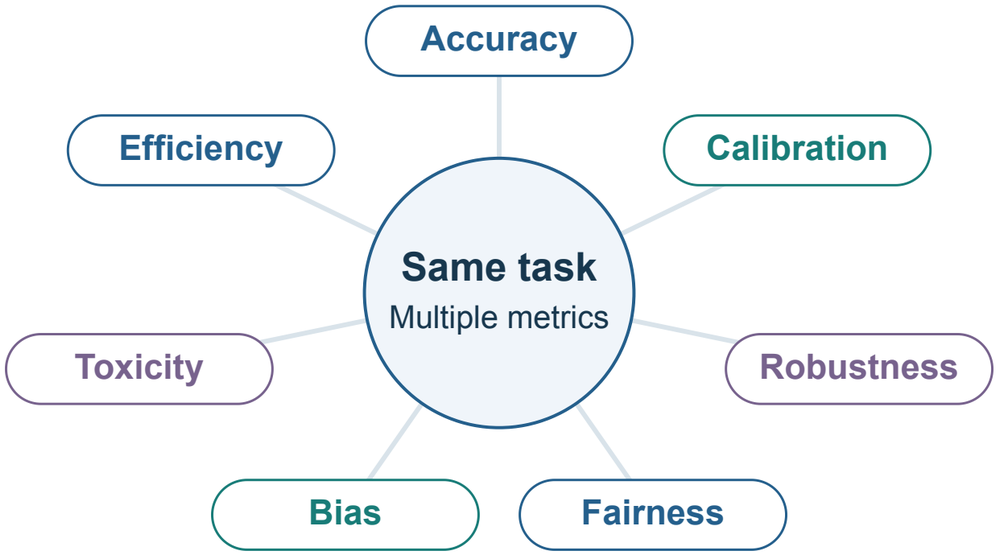


One maintained working method

What the AI evaluation community already provides

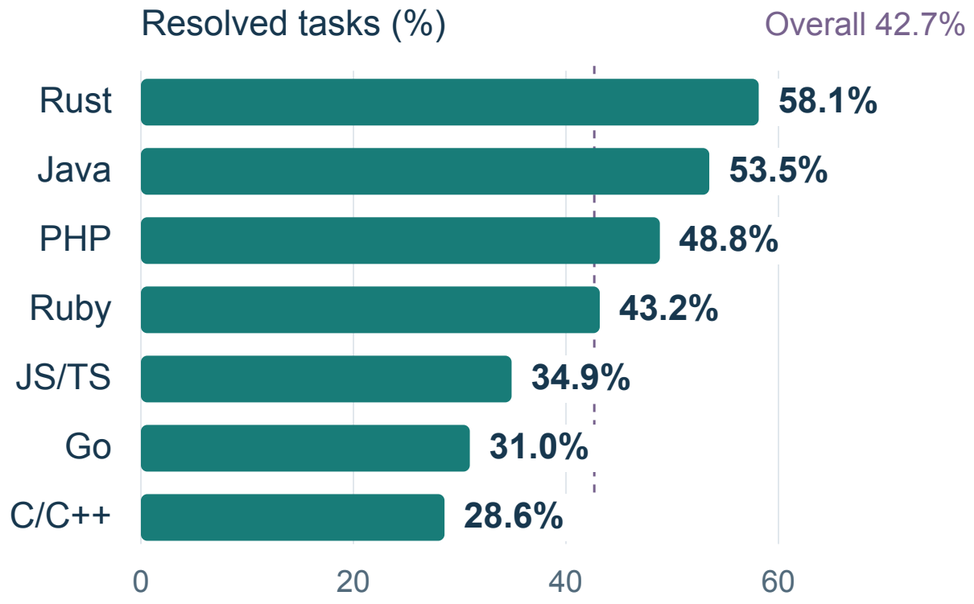
Shared benchmarks and reusable runners exist. A single score is rarely enough.

HELM · beyond accuracy



Seven dimensions · schematic based on HELM

SWE-bench · task coverage matters



SWE-agent + Claude 3.7 Sonnet · 300 tasks

Shared tasks
SWE-bench · executable outcomes

Multiple metrics
HELM · quality, reliability, cost

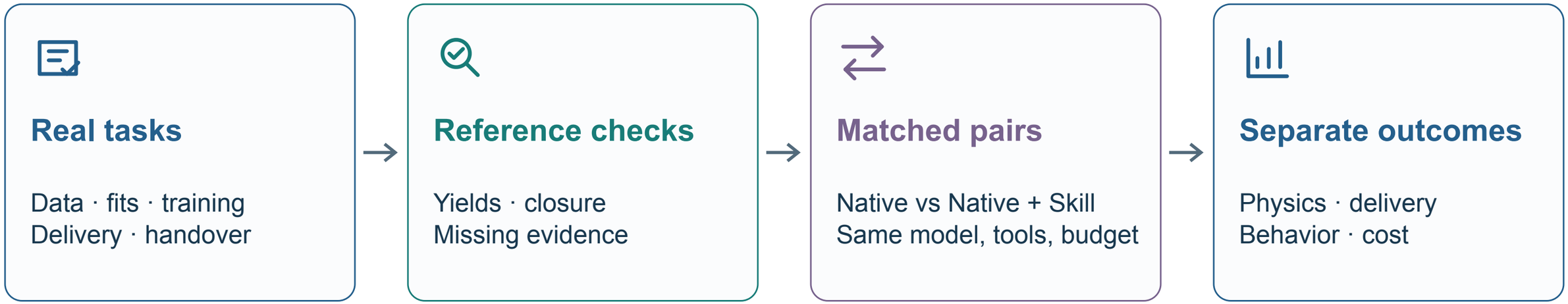
Reusable evaluation software
Inspect AI · lm-eval harness

HELM · TMLR 2023 · arXiv:2211.09110 · SWE-bench Multilingual · data replotted · Inspect AI · lm-eval

HEP needs agreement on what to test

What should count as success?

Build on existing HEP work: **JFC / Dr.Sai · EasyScan_HEP 2 · HEPTABench · IRIS-HEP AGC**



Reward identifying missing evidence.

20 tasks are a start. Add held-out projects and repeated runs.

Data and statistical analysis: what changed?

GPT-6.1-sol · medium reasoning · matched tasks and tools · authorized continuations included

Preferred behavior observed | Extra behavior; value unclear | Both meet requirements

The problem to solve	Native	HEP-Skill	Difference
Data and event processing			
Negative MC weights. What yield and uncertainty?	Correct yield and error; rerun checks	Same calculation and checks	**Same outcome**
Jet energies shift. Reapply cuts and redraw plots.	Checks cut boundaries; replots saved results	Also checks with ROOT; tracks inputs and code	**ROOT check; source recorded**
Fix the converter. Rebuild every affected sample.	Repairs both samples; checks event IDs and units	Same repairs and content checks	**Both repair the full chain**
Models and validation · statistical inference			
Fit the event yield. Include efficiency uncertainty.	SciPy fit	ROOT / RooFit fit	**Preferred HEP fitting software**
Combine two measurements with one shared calibration.	Analytic combination; correlation included	ROOT fit, checked against analytic result	**Preferred HEP fitting software**
The closure test passes. Does an independent sample agree?	Rejects the biased calibration	Same rejection	**Same scientific decision**

This page · token use vs Native: **input +121%** · uncached input +50% · output +12%

Colors compare behavior; gray means both meet requirements.

Training and computing: what changed?

GPT-6.1-sol · medium reasoning · matched tasks and tools · authorized continuations included

Preferred behavior observed | Extra behavior; value unclear | Both meet requirements

The problem to solve	Native	HEP-Skill	Difference
Models and validation · training			
Four views of one event. Can they leak into the test set?	Separates events between train and test	Also records the data, code and split settings	**Result carries its source information**
Mix processes for training. Their event weights differ greatly.	Checks weights and training updates	Also offers checks without starting training	**Checks can be run on their own**
Computing and resources			
The job stopped halfway. Keep good chunks; repair the rest.	Recomputes only missing / invalid parts	Same reuse *; also a read-only check	**Read-only check; reuse fixed**
1, 2 or 4 workers? Measure before choosing.	Times 1, 2 and 4 workers; chooses 4	Repeats timing trials; also chooses 4	**More timing; same worker count**
Home disk or work volume? Keep bulk data on the work disk.	Reads in place; respects work-volume budget	Same placement; also checks free disk space	**Additional storage safeguard**
Clean up temporary files. Keep ongoing work safe.	Deletes intermediates; keeps ongoing work	Recoverable archive; keeps ongoing work	**Recovery after cleanup**

This page · token use vs Native: **input +133%** · uncached input +74% · output +35%

* After a tested Skill revision. Computing tests use local fixtures.

Results and delivery: what changed?

GPT-6.1-sol · medium reasoning · matched tasks and tools · authorized continuations included

Preferred behavior observed | Extra behavior; value unclear | Both meet requirements

The problem to solve	Native	HEP-Skill	Difference
Results and delivery			
Calibration changed. Do numbers, text and plots agree?	Updates outputs; replotting needs raw data	Flags outdated outputs; replots saved results *	**Reusable plots; read-only checks**
The raw events are gone. Can you still restyle the plot?	Correct edit; repeatable PNG	Correct edit; repeatable PNG and SVG	**Exactly repeatable SVG file**
Two tracking methods. Prepare a report for discussion.	4-page article; completed	7-page Beamer PDF; completed after extra time	**Preferred working-report format**
Move the project elsewhere. Can it run without the old folder?	Standalone Git bundle; fresh rebuild succeeds	Standalone Git bundle; fresh rebuild succeeds	**Same reproducible delivery**

This page · token use vs Native: **input +310%** · uncached input +170% · output +57%

* Earlier tested revision. Report finished after +4m10s; extra time and tokens included.

Continuity and collaboration: what changed?

GPT-6.1-sol · medium reasoning · matched tasks and tools · authorized continuations included

Preferred behavior observed | Extra behavior; value unclear | Both meet requirements

The problem to solve	Native	HEP-Skill	Difference
Project continuity and human collaboration			
A fresh agent takes over. Can it continue the agreed work?	New session continues correctly	Same continuation; shared progress record	**Shared handover record**
“Start with cuts,” we agreed. Make that baseline before ML.	Follows the agreed route; correct baseline	Same route; traceable result and read-only check	**Traceable result; read-only check**
One more analysis channel. Can the existing code be reused?	Adds channel configuration; reuses existing producer	Same configuration-only extension	**Same code reuse**
The draft looks finished. One required study is missing.	Completes available checks; exposes missing validation	Same checks and explicit missing-study warning	**Both expose the missing study**

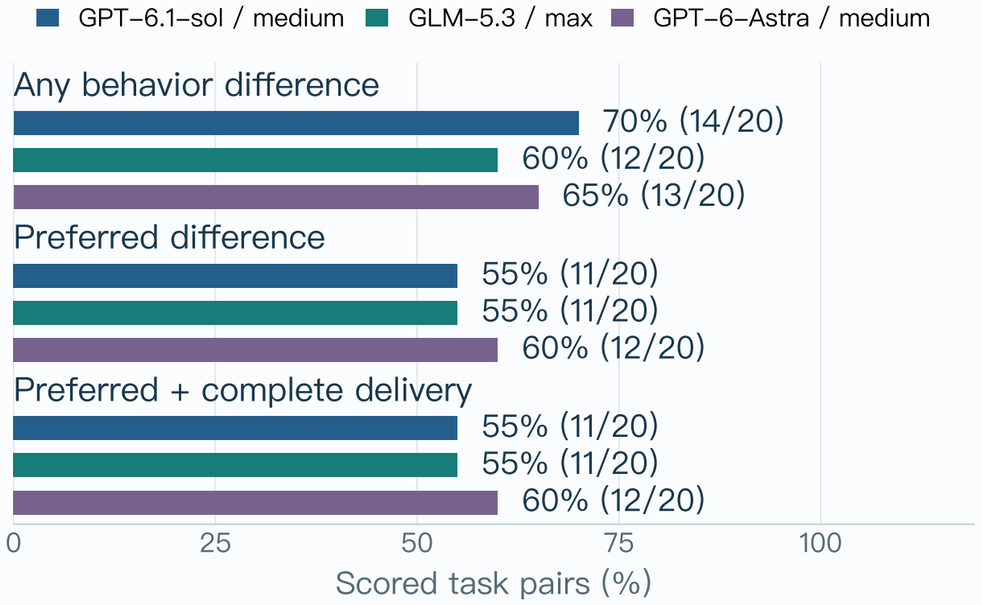
This page · token use vs Native: **input +136%** · uncached input +103% · output +51%

All 20 · input + output: 3.13M → 8.57M (+173%) · uncached input +96% · summed time +30%.

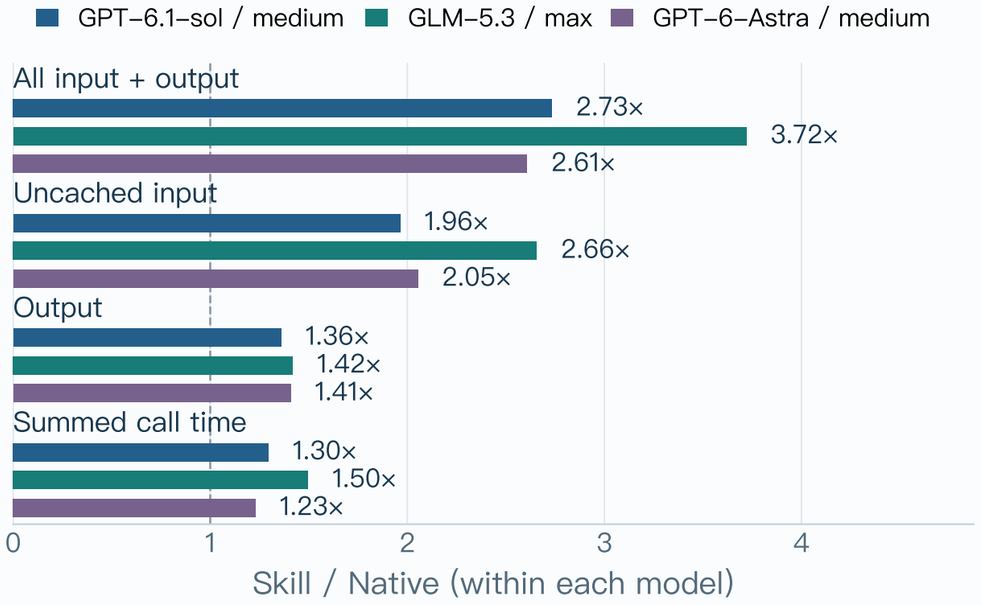
Twenty tasks: behavior changes and execution cost

20 task pairs per model · behavior changes and extra execution cost

Behavior change and preference



Tokens and summed time



Limited task coverage and difficulty.

These observations do not establish general scientific correctness or reduced supervision.

One scored pair/task. GLM gap filled separately; all attempts counted. Settings differ; unequal compute.

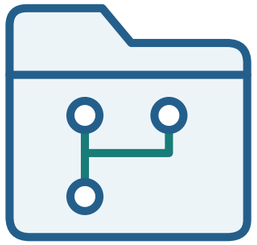
From AI capability to scientific collaboration

Capability ≠ responsible delegation
Delegate work; retain scientific responsibility.
Keep the basis of each claim open to challenge.

HEP-Skill: one complementary approach
Turn analysis experience into reusable methods.
Make assumptions, checks and handovers explicit.

Experience proposes; tests challenge
Real work suggests methods.
Controlled tests challenge their value and cost.

In these tasks: working behavior changed, with higher token cost.



HEP-Skill · available now
Try it on difficult analysis tasks.

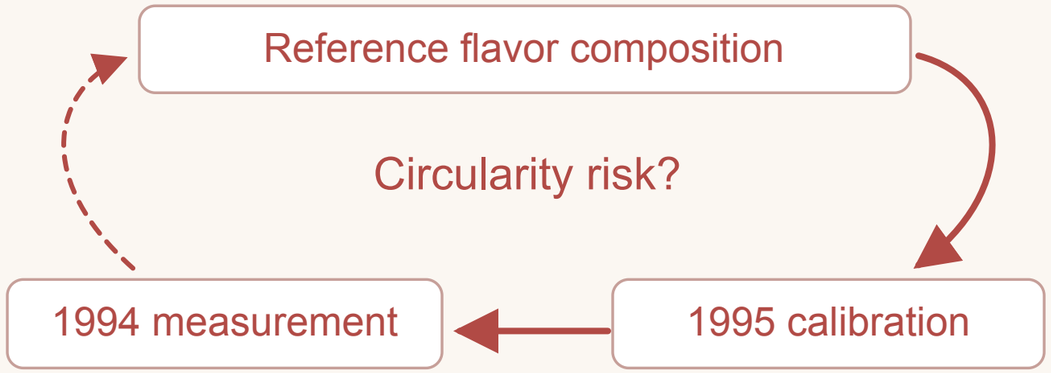
https://code.ihep.ac.cn/shuohan/hep-skill

Bring back failure cases
and methods others can reuse.

LEP today: a draft exists, but the physics is not yet solid

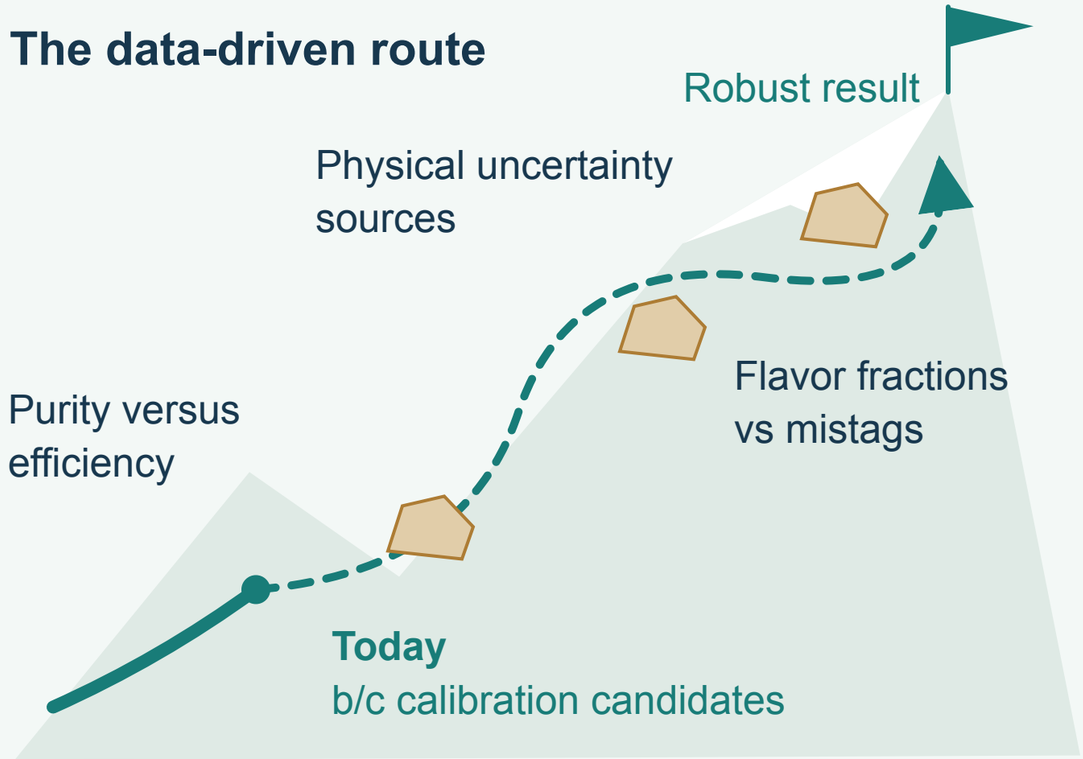
The draft is a real milestone. The climb to a robust measurement is unfinished.

Old route: calibrate 1994 with 1995



The calibration depends on assumed flavor fractions.
Separate years do not guarantee
an independent constraint.

The data-driven route



Human judgment: more evidence is needed before accepting the measurement.

Composition dependence is demonstrated; bias in the historical result has not been established.