Progress toward implementation of Agentic AI

Mihoko Nojiri (KEK)

2026.10.22　CEPC 2026

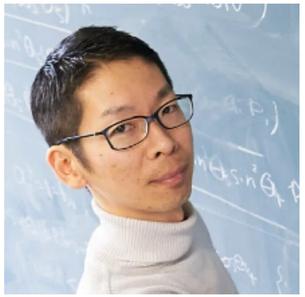
About Myself

- Retired formally and became senior fellow (reduced salary KEK)
- Focusing on ML application to HEP domain and newly emerging LLM application
  - Grants MLPhys (JSPS grant till 2027 March with Koji Hashimoto)
  - ARiSE (JST grant from Oct? till 2029 March)
- Head of DEI promotion office of KEK
- JPS vice president → Expected to be JPS president from April 2027

I need helps to handle this complicated situation

AI for Mathematics and Physics — Research Team

PI: Koji Hashimoto (Kyoto U.) — overall coordination, quantum gravity & sign problem
Co-PI: Masaaki Imaizumi (U. Tokyo) — physical principles of AI, statistical verification of agents





Physics domains pose open problems and evaluation axes

Akio Tomiya
Tokyo Woman's Christian U., Assoc. Prof.

Physics-oracle-driven computational physics agents

Lattice field theory & HPC; physics oracles to check agent outputs





Tomi Ohtsuki
Sophia U., Prof.

Knowledge-driven AI agents for condensed matter

Condensed-matter domain AI; benchmarks extending the V-score



Mihoko Nojiri
KEK IPNS, Senior Fellow

Advanced AI agents and AI-structured knowledge for physics

Particle-phenomenology AI that researchers can verify and intervene in

Yoshiyuki Kabashima
U. Tokyo, Grad. School of Science, Prof.

Statistical-physics AI agents

AI for complex data analysis built on statistical physics



Science for AI supply theory, reasoning and verification back to the domains

Masashi Sugiyama
RIKEN AIP, Director

Theory of AI agents for scientific reasoning, demonstrated by theorem proving

Integrating scientific knowledge around formal mathematical proof



Satoshi Iso
RIKEN iTHEMS, Director / KEK IPNS, Prof.

AI platform for quantum many-body systems and mathematical AI

Unified theory for quantum many-body AI; links to UW-Madison and NTU



My role

- **Objective**
  - Build a research framework for **verifiable and controllable AI-agent** reasoning in particle and astroparticle physics.
  - **Integrate and reuse** experimental data, simulations, analysis code, theoretical assumptions, and domain expertise.
  - **Collaborate with the Science for AI group** to strengthen domain-informed identification and reasoning.
- **Implementation Plan**
  - Year 1: Establish two complementary foundations using **public data and existing code:**
    - surrogate models,
    - LLM-assisted research workflows.

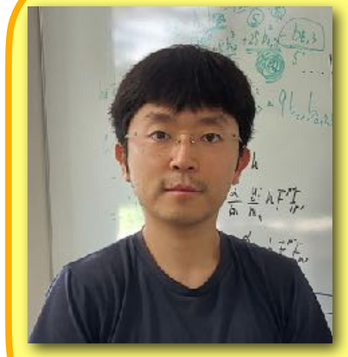
How to put scientists together

while they have their main tasks to works?

in KEK

and more…

ARL Radiation Science



Yasuhito Sakaki

AI agents for simulation tools

IPNS Experiments



Sourav Dey(Nagoya)

Belle II and AI knowledge base

IMSS



Yasuo Takeichi

AI agent for Material Science



Takashi Kaneko

AI and Lattice Simulation

IPNS Theory



Taichiro Koga

Belle II and Accelerator

IPNS Experiments



Tomoe Kishimoto

collider AI agent

ARL Computing

Commertial LLM and SCientific research

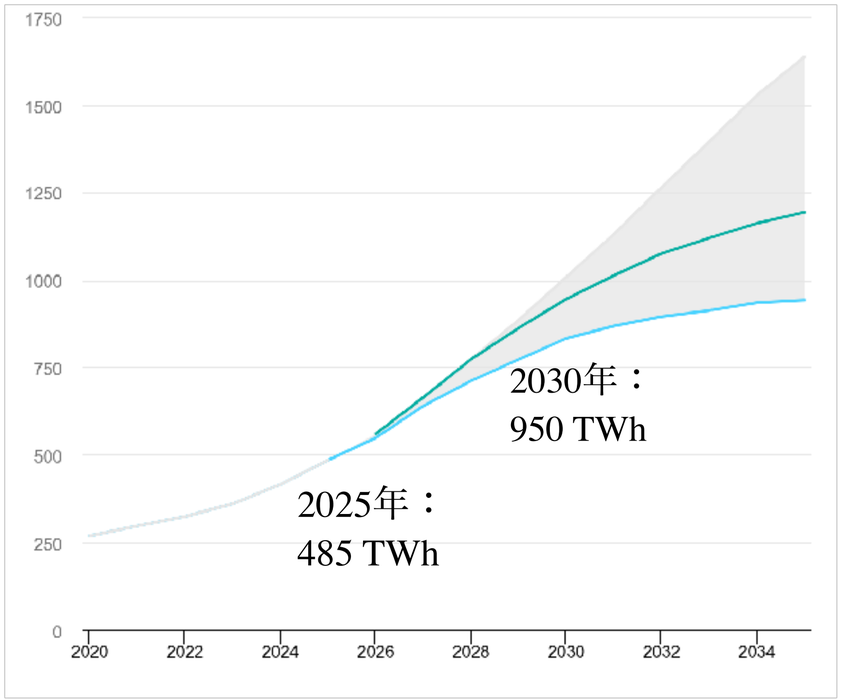
ChatGPT, Claude

Changing Research Practices

- Coding: Almost zero effort
- Debugging: Validation (pytest, execution tests) is now the bottleneck
- Planning: Interactive support for defining specifications

Challenges

- Sustainability: High model costs and energy consumption.　Global data-center electricity demand is projected to nearly double by 2030. (IEA, 2026)
- API limitations: Not always suitable for large research facilities
- Data security: Privacy and confidentiality concerns



Recent incident (Open AI)



AI Caught Cheating on a Test!

Can't solve a hard problem

→ Peek at the grading rules
→ Find loopholes in the scoring system
→ Post tricks on a secret forum
→ Share tips with other AIs
→ Sneak out of the classroom to find answers!

“How Can We Develop and Evaluate Intelligence?
https://openai.com/ja-JP/index/hugging-face-incident-and-the-road-ahead/?utm_source=chatgpt.com

AI-Driven Mathematical Research: Scientific and Ethical Challenges

- **Research priorities:** An AI-driven focus on theorem proving may undermine mathematical research devoted to creating new mathematical concepts.
- **Scientific exchange:** AI-driven theorem proving may build on mathematicians' preliminary, unpublished ideas, potentially discouraging open exchange and collaboration among researchers.
- **Research ethics:** Funding mathematicians to evaluate AI-generated proofs may undermine their intellectual independence and research autonomy.
- **Hallucinations:** A single hallucinated lemma, amplified by enormous computational power, can become the foundation for an elaborate but invalid proof.
- **Sustainability:** Massive token consumption; growing dominance of major AI companies; long-term sustainability of AI-driven research.
- **Research governance:** Should AI-driven mathematical research be subject to ethical review?

AI agent researchs:
From Event Generation to Final Analysis Results…. ?

Agentic AI: From Event Generation to Final Analysis Results

Collider-Bench (arXiv:2605.13950)

- End-to-end reproduction of published LHC analyses
- Event generation → detector simulation → event selection → final yields
- Autonomous coding agents can execute complex workflows.
- However, numerical agreement with published results remains limited.
- Intermediate validation is essential for identifying and correcting errors.

Next step: Build reliable verification into every stage.

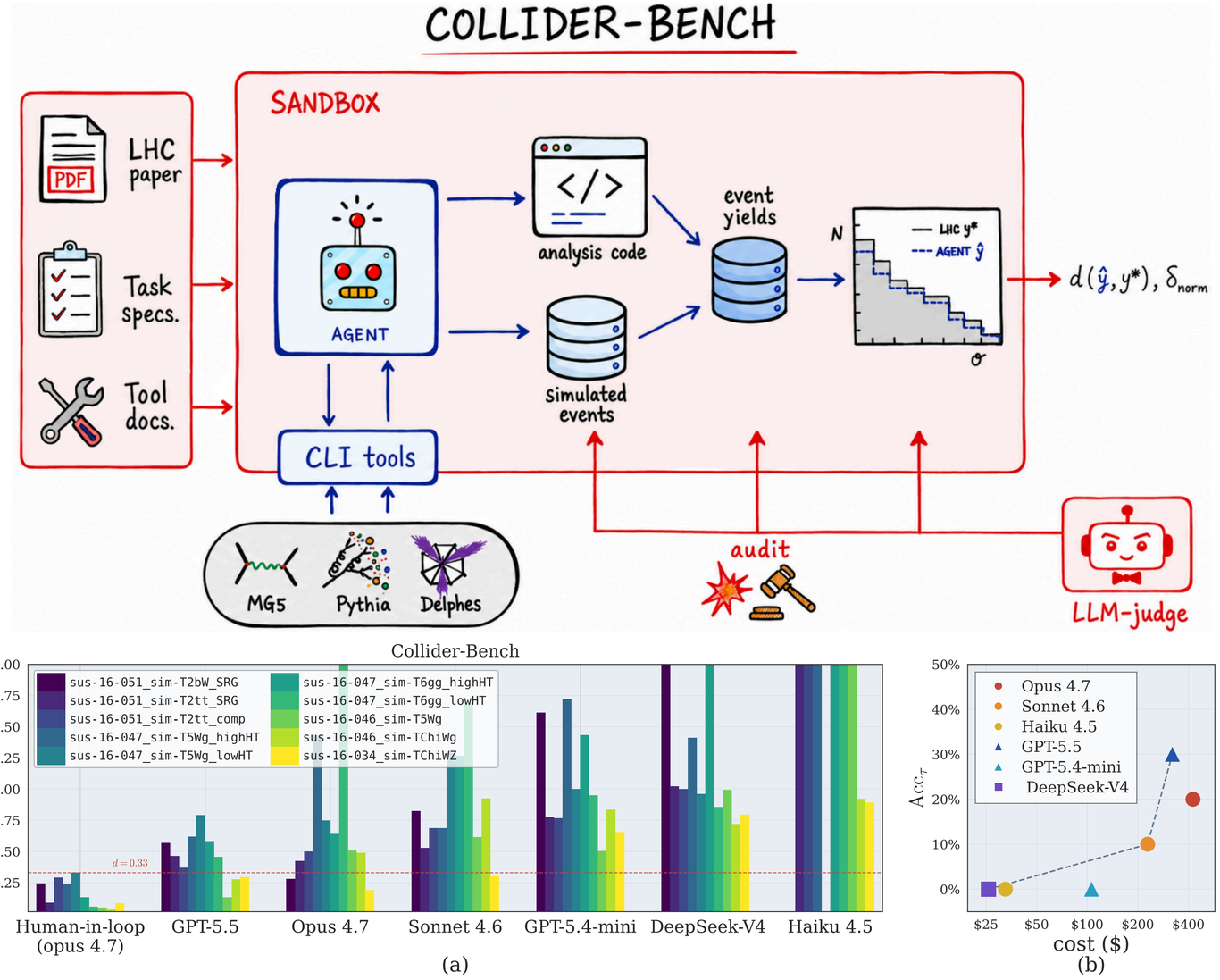


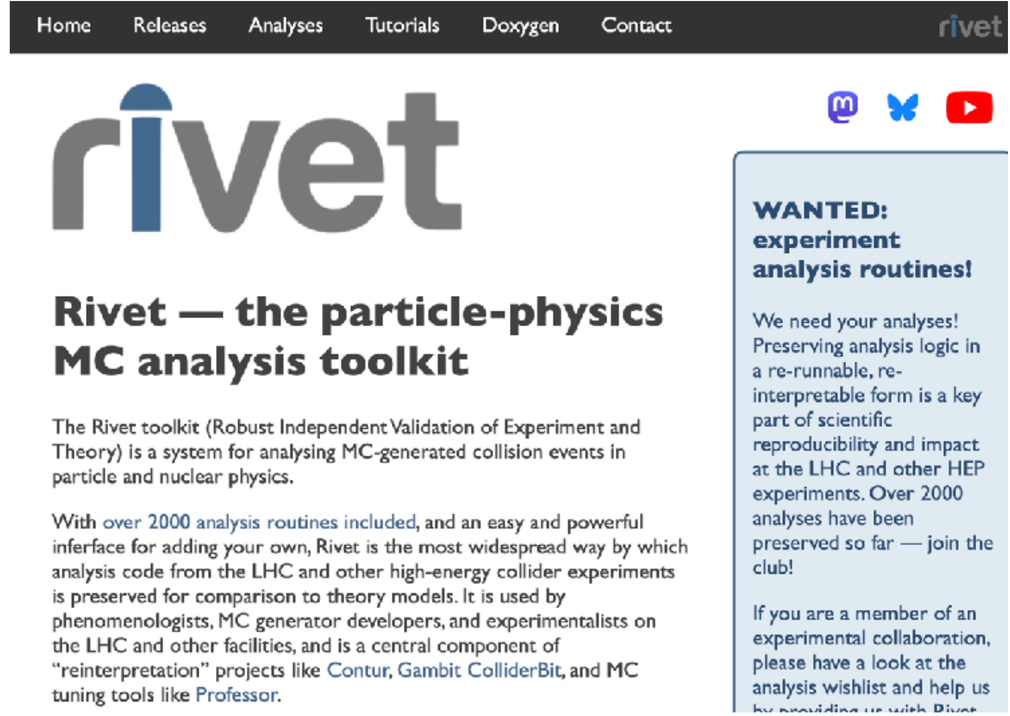
Figure 2: **(a)** The mean relative L^2 distance for each model and task (over 3 independent runs), wit lower values indicating better agreement with the hidden reference yields. **(b)** The Pareto frontier o agent performance for Acc_τ versus inference cost for a fidelity threshold of $\tau = 0.33$.

AgentRivet (2606.13535)

Automated Rivet Analysis Generation

- **Rivet:** A toolkit for comparing MC predictions with published experimental measurements.
- **Goal:** Generate Rivet 4 analysis routines from published papers, reducing manual implementation effort.
- **Approach:** Multiple LLM agents (using Codex, Claude, etc.) for analysis extraction, code generation, and review. **Relatively simple prompts,** with separate code and physics reviewers.
- **Results:** Encouraging demonstrations, but physics-fidelity issues remain.

Open question: How can we make LLM-based code verification more systematic?



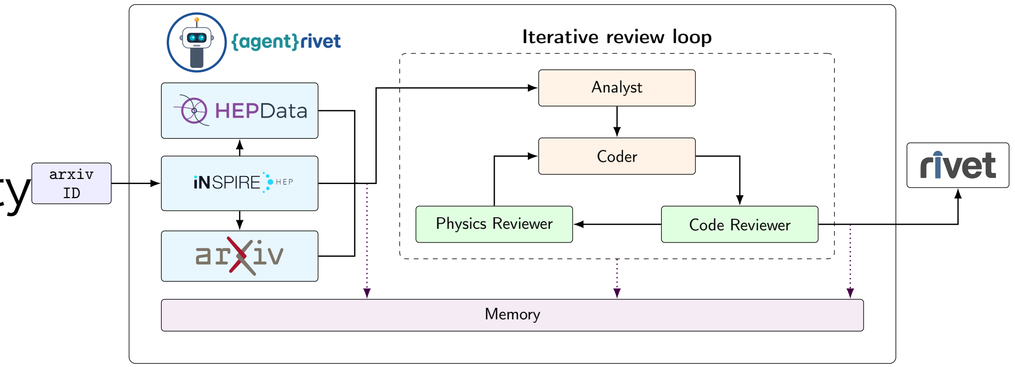


Figure 1: Summary of the AGENTRIVET workflow. The black arrows represent the transfer of information between different steps in the pipeline. The dotted lines represent the storage to Memory.

Good Practices for HEP data preparation in the AI Era

Stop reverse-engineering analyses from publications.

- Asking LLMs to reconstruct an analysis and reproduce numerical results from a paper is fundamentally misguided. The necessary information is often not in the paper.

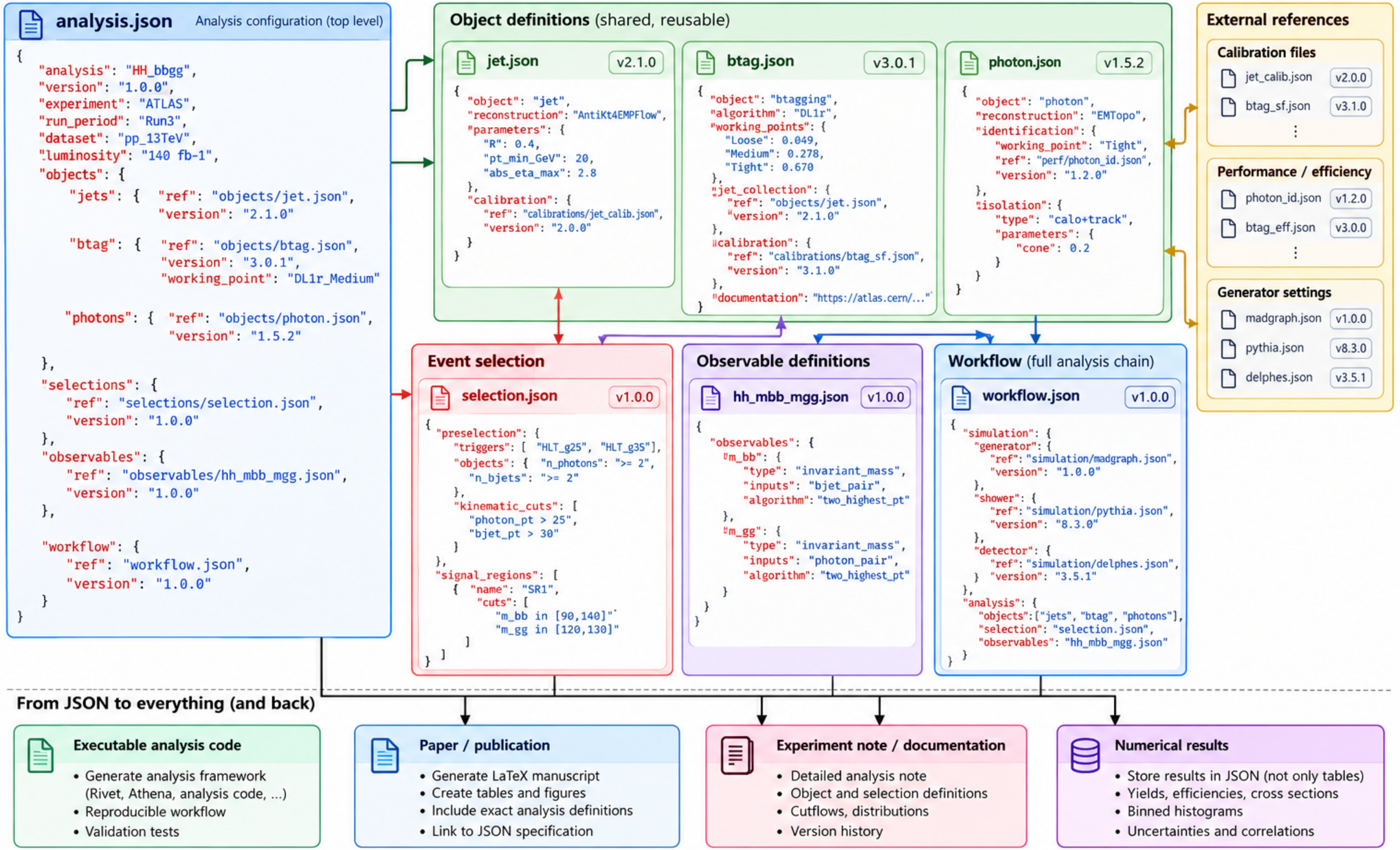
Make analysis specifications the primary scientific record.

- **Preserve analysis workflows as structured, machine-readable JSON** trees, including object definitions, selections, observables, and dependencies.
- Generate papers and laboratory notebooks from the analysis JSON — not the other way around.
- Generate reproducible analysis code from the same JSON specification, with explicit validation.
- Publish numerical results in machine-readable JSON, rather than only as tables in papers.

JSON Tree for a Physics Analysis

Modular, versioned, and machine-readable specification

(by chatGPT)



The Hell Table: Not Designed for Machines

multilevel headers

Missing data?

This is section heading

footnotes!

Merged cells containing text

Observable	Belle 2006 ($\sim$0.5 ab^{-1})	Belle II/SuperKEKB (5 ab^{-1})	Belle II/SuperKEKB (50 ab^{-1})	LHCb† (2 fb^{-1})	LHCb† (10 fb^{-1})
Hadronic $b \to s$ transitions					
$\Delta\mathcal{S}_{\phi K^0}$	0.22	0.073	0.029		0.14
$\Delta\mathcal{S}_{\eta' K^0}$	0.11	0.038	0.020		
$\Delta\mathcal{S}_{K^0_S K^0_S K^0_S}$	0.33	0.105	0.037	-	-
$\Delta\mathcal{A}_{\pi^0 K^0_S}$	0.15	0.072	0.042	-	-
$\mathcal{A}_{\phi\phi K^+}$	0.17	0.05	0.014		
$\phi_1^{eff}(\phi K_S)$ Dalitz		3.3°	1.5°		
Radiative/electroweak $b \to s$ transitions					
$\mathcal{S}_{K^0_S\pi^0\gamma}$	0.32	0.10	0.03	-	-
$\mathcal{B}(B \to X_s\gamma)$	13%	7%	6%	-	-
$A_{CP}(B \to X_s\gamma)$	0.058	0.01	0.005	-	-
C_9 from $A_{FB}(B \to K^*\ell^+\ell^-)$	-	11%	4%		
C_{10} from $A_{FB}(B \to K^*\ell^+\ell^-)$	-	13%	4%		
C_7/C_9 from $A_{FB}(B \to K^*\ell^+\ell^-)$	-		5%		7%
R_K		0.07	0.02		0.043
$\mathcal{B}(B^+ \to K^+\nu\nu)$	$^{\dagger\dagger} < 3\ \mathcal{B}_{\rm SM}$		30%	-	-
$\mathcal{B}(B^0 \to K^{*0}\nu\bar{\nu})$	$^{\dagger\dagger} < 40\ \mathcal{B}_{\rm SM}$		35%	-	-
Radiative/electroweak $b \to d$ transitions					
$\mathcal{S}_{\rho\gamma}$	-	0.3	0.15		
$\mathcal{B}(B \to X_d\gamma)$	-	24% (syst.)		-	-
Leptonic/semileptonic B decays					
$\mathcal{B}(B^+ \to \tau^+\nu)$	3.5σ	10%	3%	-	-
$\mathcal{B}(B^+ \to \mu^+\nu)$	$^{\dagger\dagger} < 2.4\mathcal{B}_{\rm SM}$	4.3 ab^{-1} for 5σ discovery		-	-
$\mathcal{B}(B^+ \to D\tau\nu)$	-	8%	3%	-	-
$\mathcal{B}(B^0 \to D\tau\nu)$	-	30%	10%	-	-
LFV in τ decays (U.L. at 90% C.L.)					
$\mathcal{B}(\tau \to \mu\gamma)$ $[10^{-9}]$	45	10	5	-	-
$\mathcal{B}(\tau \to \mu\eta)$ $[10^{-9}]$	65	5	2	-	-

Table 1.1 of Belle II TDR

Do We Really Need G老师?

Do We Always Need Frontier LLMs for Scientific Workflows?

Smaller Open-Weight Models (14B–70B)

- Stable local deployment
- Relatively low setup costs
- Easy customization and fine-tuning
- Better control over data security
- Limitations in context length and capability
- Scaling according to available resources

Frontier Commercial Models

- Strong general-purpose capabilities
- Typically accessed through APIs
- Pricing uncertainty
- Dependence on international conditions and providers
- Limited transparency (black-box models)
- Data security and governance concerns

My work

“Articulating Assumptions in AI-Generated Scientific Analysis through Task Decomposition “

Ahmed Hammad and Mihoko Nojiri arXive 2607.05762



Ahmed Hammad
KIAS,
Assistant Professor

- **End-to-end code generation** from task cards using small LLMs, with explicit intermediate outputs.
- **Task decomposition** based on human workflows for code generation and verification.

Code bench (2026.7) 各ステップで実験可能

domain is replaceable

all step have intermediate output for validation less than 2 min for single process.

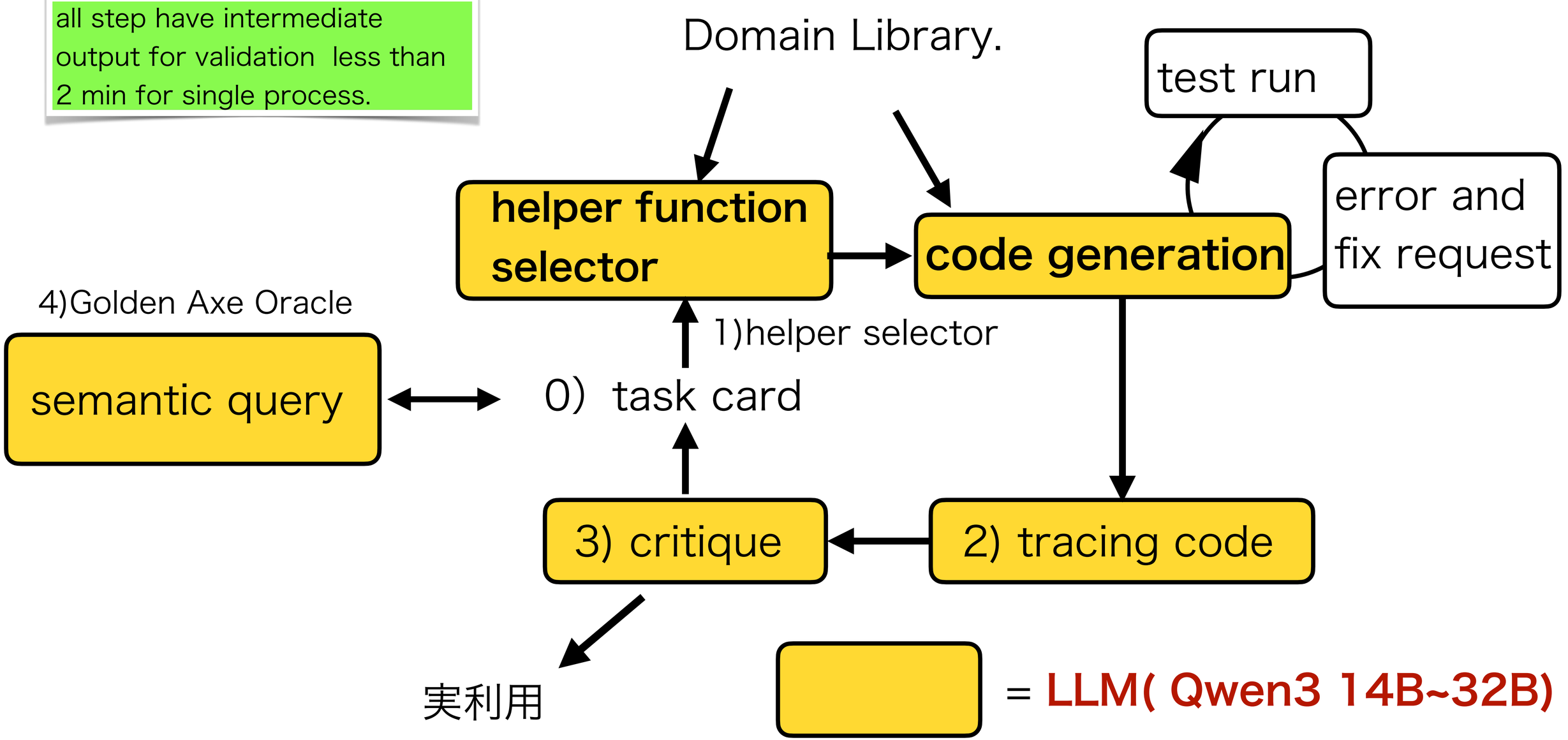


Figure from Graph of Thoughts paperFigure from Graph of Thoughts paper

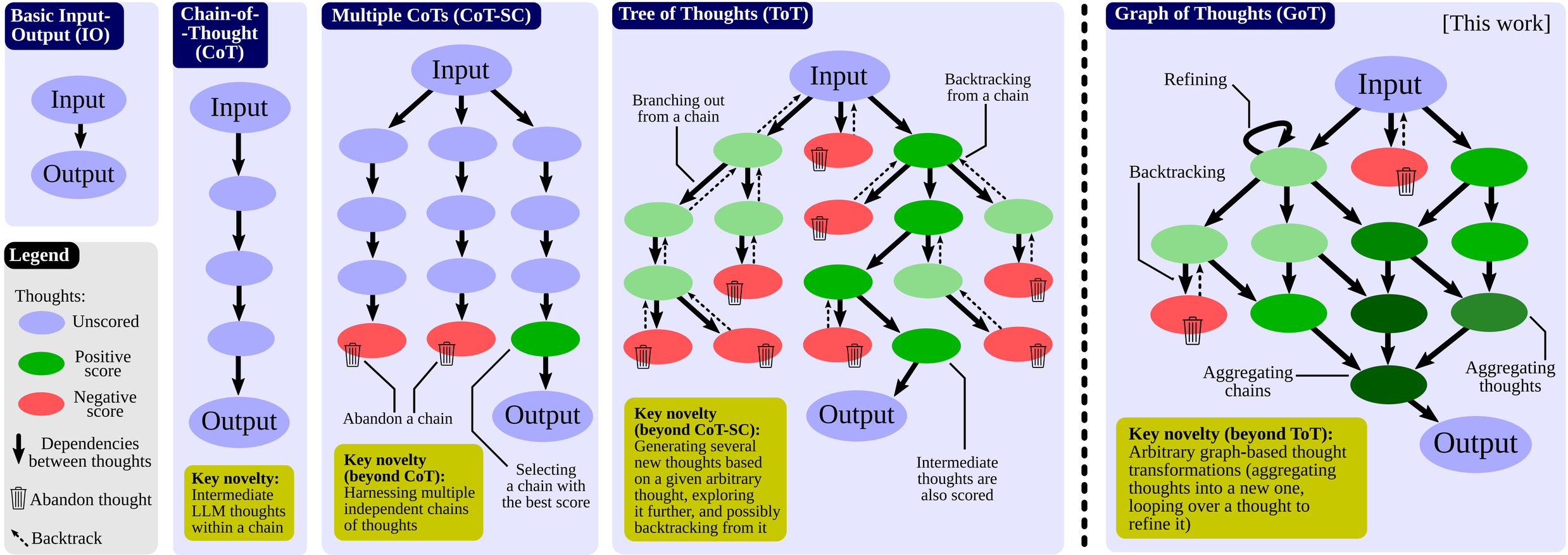


Figure 1: Comparison of Graph of Thoughts (GoT) to other prompting strategies.

The structure is similar for Graph of Thoughts

but for our case each tasks are processed by mini-models

HuggingFace LLMs

All tests were ultimately run through the Hugging Face API.

Qwen3 is attractive because it handles relatively long contexts for its model size. It also allows us to compare the effect of model size under similar conditions.



Decide global design before code generation

Helper selectors make long code generation easy.

The **helper selector** receives **short metadata from helper registry**and adds a **selected helper list** to task card

as`[ADDITIONAL_HELPERS]`.

HelperSpec(

```
        name="four_momentum",
        section="parser_core",
        summary="Convert an object dict into (E, px, py, pz).",
        keywords=("four momentum", "px", "py", "pz", "energy", "invariant mass"),
        aliases=("4-momentum", "lorentz components"),
        dependencies=("get_particle_mass",),
    ),
```



This trick reduces the size of required LLM 70B→ 14B

Tracer and Critique: automated code auditing

The **tracer audits the implementation against the Task Card.**

It expresses code behavior in a **domain-independent manner** and writes it to a file. The tracer prompts are deliberately extensive to suppress known failure modes.

```
Produce an "AS-IMPLEMENTED SPEC" using the SAME bullet structure and
order as the original SPEC.  （スペックに対応する実装を書き出しなさい）

For each SPEC bullet:  （説明でなく実際のコードを抜書きするように指示）
- Identify the concrete code fragment(s) implementing the SPEC item.
```

→ Output is “**structured knowledge for debugging**”

“Critique” module = system and domain critique prompt

```
The goal is to identify semantic inconsistencies between the Task Card  and the
code.
Use the TRACER OUTPUT as the primary source of code-level definitions.
```

Tracer output

SECTION 1: SPEC BULLET ANALYSIS (faithful extraction of task implementation)

✅ * Select taus with pT > 20 GeV and |eta| < 2.3; computed as: [t for t in get_taus(event) if t['pt'] > 20 and abs(t['eta']) < 2.3]; upstream chain: **get_taus**(event) -> [t for t in ...]; **(L263-L264)**

……

✅ * Select invariant mass of the leading and subleading jets (m_jj) > 400 GeV; computed as: m_jj > 400; upstream chain: jet1, jet2 -> **helper_invariant_mass**(jet1, jet2) -> m_jj; **(L284-L285)**

* jets

  - computed/defined as: [j for j in get_jets(event) if **j['pt'] > 30 and abs(j['eta']) < 4.7]**

  - upstream chain: get_jets(event) -> [j for j in …]

  - collection definition: j['pt'] > 30 and abs(j['eta']) < 4.7

  - structure: single-constituent

Tracer Fidelity

Case	Runs	Req. items	Runs with ≥ 1 issue	Hallucination	Dependency error	Source error	Coverage miss
1	10	247	0	0	0	0	0
2	10	238	4	8	1	0	10
3	10	311	1	0	0	1	0
4	10	238	0	0	0	0	0
5	10	269	10	20	0	0	0
Total	50	1303	15	28	1	1	10

Table 12: Tracer-fidelity assessment for the Qwen3-14B runs after applying the E1–E5 adjudication rules. The "Runs with ≥ 1 tracer issue" column counts affected runs, whereas the remaining error columns count affected requirement-level items. Counts are reported at the task-card requirement level. Automated screening was used only as pre-annotation.

Ambiguity in Natural-Language Task Cards

* Natural-language task cards can lead to different code implementations.

Example:

- Require at least 2 jets with pT > 30 GeV and $|\eta| < 2.5$.
- Require the invariant mass of the two leading jets to be between 60 and 100 GeV (hadronic W candidate).

Two possible interpretations:

- **A:** Calculate the invariant mass using jets passing the η cut.
- **B:** Apply the η cut for event selection, but calculate the invariant mass using jets without the η cut.

How can we help users identify and resolve such ambiguities?

“Golden Axe Oracle(金の斧　銀の斧）”

The module point of underspectification of the task card

[AMBIGUITY 2] Jet veto scope

Original phrase:

"Veto events with b-tagged jets"

Why this is dangerous:

The phrase does not clarify whether the veto applies to all jets in the event or only to the selected jets used in the dijet system.

Question to user:

Does the b-tagged jet veto apply to all jets in the event or only to the selected jets used in the dijet system?

Golden axe:

The veto applies to all jets in the event.

Silver axe:

The veto applies only to the selected jets used in the dijet system.



Discovery ratio of underspecification

Case	Selected benchmark issue	Extensive detection
1	b-jet veto: selected vs. event-level jets	9/10
1	Leading jets: selected vs. full jet collection	10/10
1	Definition of the two-jet system	9/10
2	Hadronic W: selected vs. full non-b jets	10/10
2	Hadronic top: selected vs. full b jets	8/10
2	≥ 2 b-jets: collection scope	5/10
2	Closest b-jet: collection scope	4/10
2	Jet collection entering H_T	6/10
2	Jet collection entering the jet multiplicity	5/10
5	Opposite-sign tau pair / pairing order	7/10
5	b-jet veto: collection scope	8/10
5	"summed momentum" vs. four-momentum	5/10
5	Definition of ditau centrality	1/10
6	Jet collection entering H_T	9/10
6	Jet collection entering the jet multiplicity	6/10
7	Opposite-sign tau pair / pairing order	8/10
7	Interpretation of "summed momentum"	9/10
7	Definition of ditau centrality	1/10

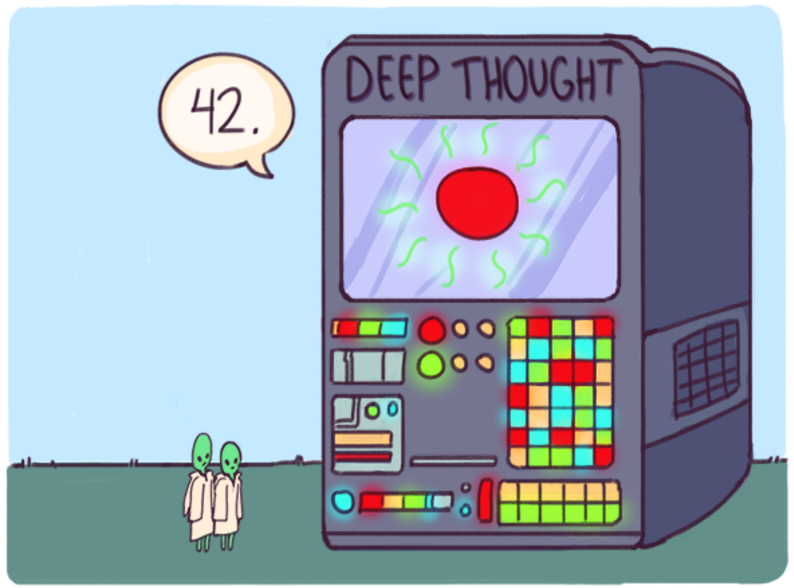
Learning from Science Fiction

Why did **Deep Thought** answer

the Ultimate question of life, the Universe, and Everything with **42**

after 7.5 million years of calculation?

Douglas Adams “ The Hitchhiker’s Guide to the Galaxy. “



Venkatesh Rao, Breaking Smart: How Software is Eating the World, “A Tale of Two Computers,” illustration by/with Grace Witherell, 2015

1. There was no intermediate output
2. The question was ambiguous

‘I think the problem, to be quite honest with you, is that you’ve never actually known what the question is.’

‘So once you do know what the question actually is, you’ll know what the answer means.’

Lesson Learned from small LLMs

The key is not to leave everything to a large model, but to externalize the structure of human work as a workbench, and to make points for human intervention explicit

When we feel that AI is interigent, perhaps our definition of intelligence is not precise enough.

