



AI-based Scientific Hypothesis Generation

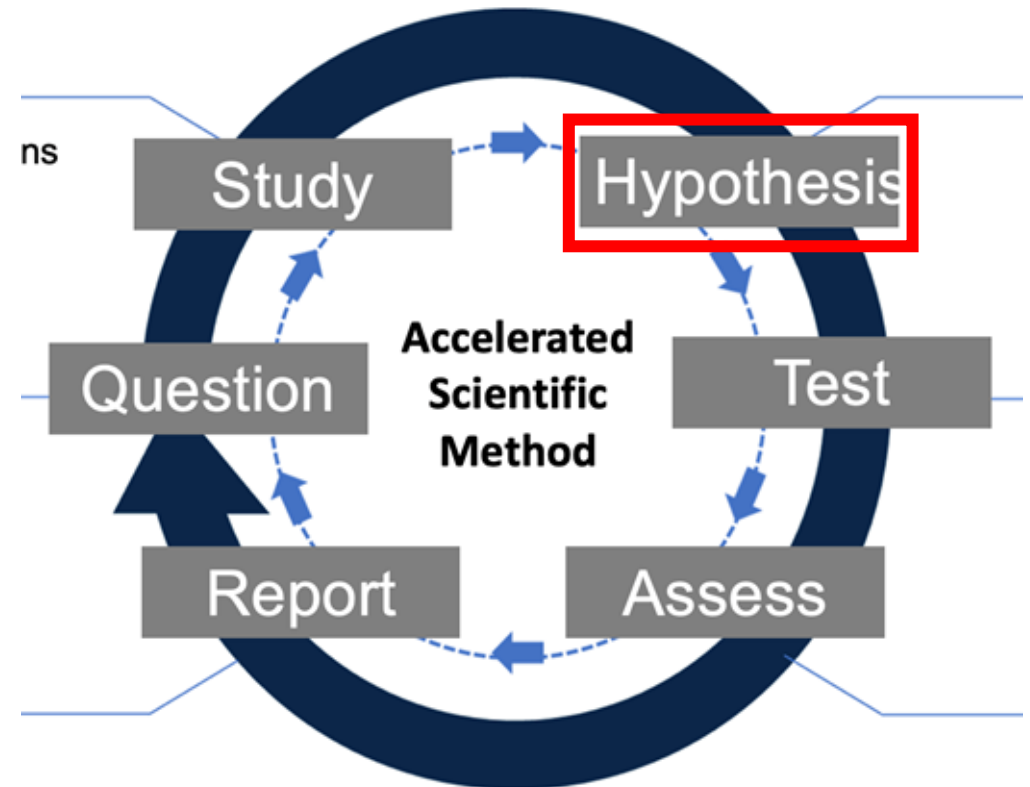
Franck Cappello

Argonne National Laboratory

EAIRA paper: <https://arxiv.org/abs/2502.20309>

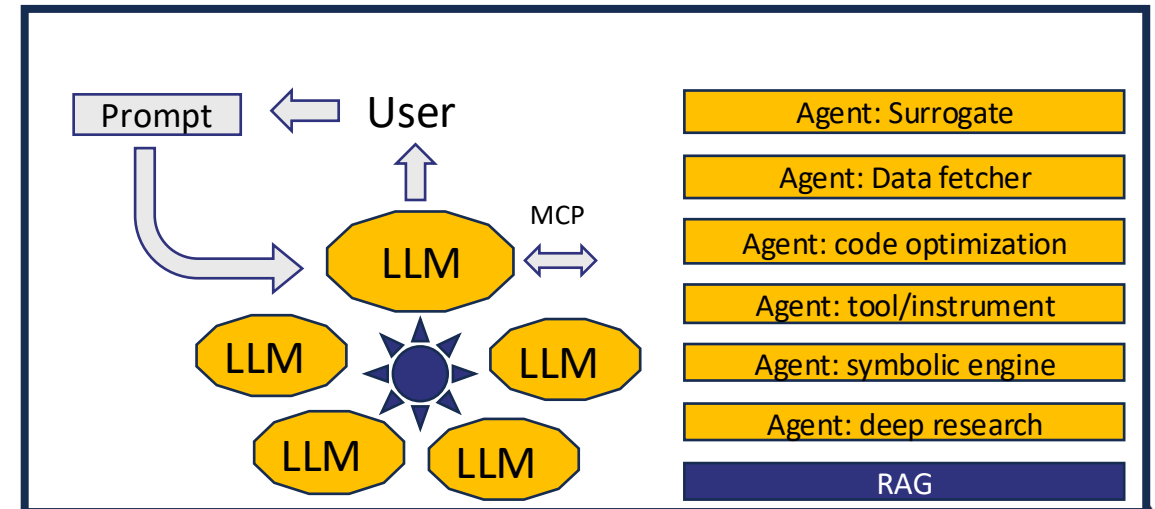
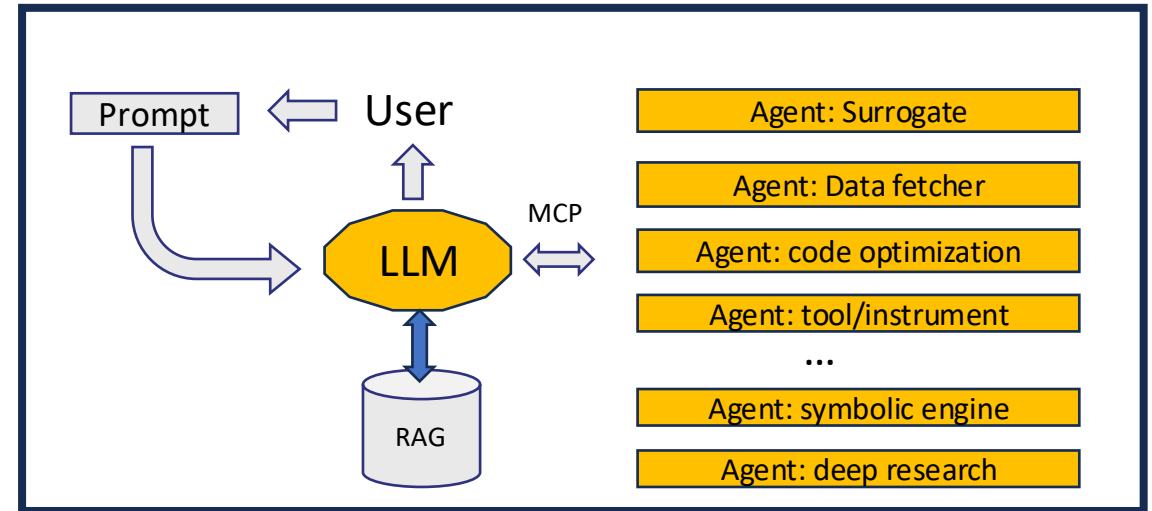
Summary

- Why Hypothesis Generation is important to consider in current road mapping efforts
- A key element is missing: deep research for scientific data
- Impact on system workload (tokens)
- Impact on HPC workload



Main Hypothesis Generation Tools (HGTs)

- LLMs augmented with agents
(Including LLMs + symbolic engines)
- Multi-modal aspect handled by agents transforming raw data in tokens/embeddings
- Co-scientists (Multi-agents – all LLMs) augmented with external agents



Generate hypotheses for what problems (list from Claude 4.6)?

- 1. Solving problems with clear specifications and constraints** Using known principles to propose solutions. e.g.: *engineers hypothesize that a new alloy composition will withstand higher temperatures based on material science theory.*
- 2. Explaining observed phenomena** Making sense of something already seen. *Example: Fleming hypothesized that a mold was producing a bacteria-killing substance after noticing clear zones around Penicillium contamination.*
- 3. Predicting future observations** Forecasting what *should* happen if a theory is correct, before anyone looks. *Example: Einstein's general relativity predicted that light would bend around massive objects — confirmed during the 1919 solar eclipse.*
- 4. Identifying causal relationships** Distinguishing correlation from causation by proposing a mechanism. e.g. *researchers hypothesize that smoking causes lung cancer via carcinogen-induced DNA mutations, not merely that the two co-occur.*
- 5. Generating new research directions** When data is sparse, hypotheses map unexplored territory. *Example: the hypothesis that gut microbiome composition affects mental health opened an entirely new field of psychobiotics research.*
- 6. Challenging or refining existing theories** Proposing that the current model is incomplete or wrong. *E.g. hypothesis that ulcers were caused by bacteria (H. pylori) directly challenged the accepted dogma that stress and acid alone were responsible.*
- 7. Unifying disparate findings** Proposing a single mechanism that accounts for seemingly unrelated observations. *Example: plate tectonics unified observations about fossils, coastline shapes, earthquakes, and volcanoes under one framework.*
- 8. Guiding experimental design** A hypothesis defines what to measure, what to control, and what would count as evidence — structuring the entire investigation.

Computer Science: The Knuth Problem

(28 Feb. 2026)

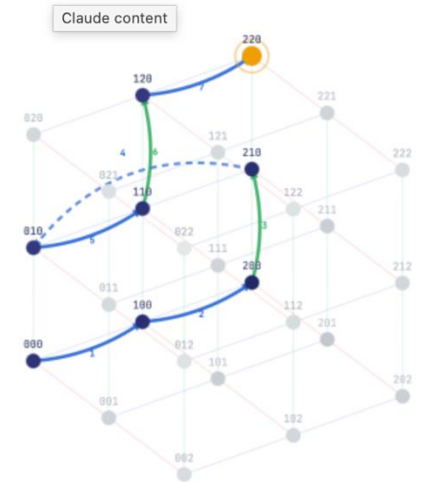
Directed m^3 -cycle

$m = 3$ · 27 vertices · all arcs go $+1 \pmod 3$ · dashed = modular wrap

Consider the digraph with m^3 vertices ijk for $0 \leq i, j, k < m$, and three arcs from each vertex, namely to i^+jk , ij^+k , and ijk^+ , where $i^+ = (i+1) \pmod m$. Try to find a general decomposition of the arcs into three directed m^3 -cycles, for all $m > 2$.

**** After EVERY exploreXX.py run, IMMEDIATELY update this file [plan.md] before doing anything else. ** No exceptions. Do not start the next exploration until the previous one is documented here.**

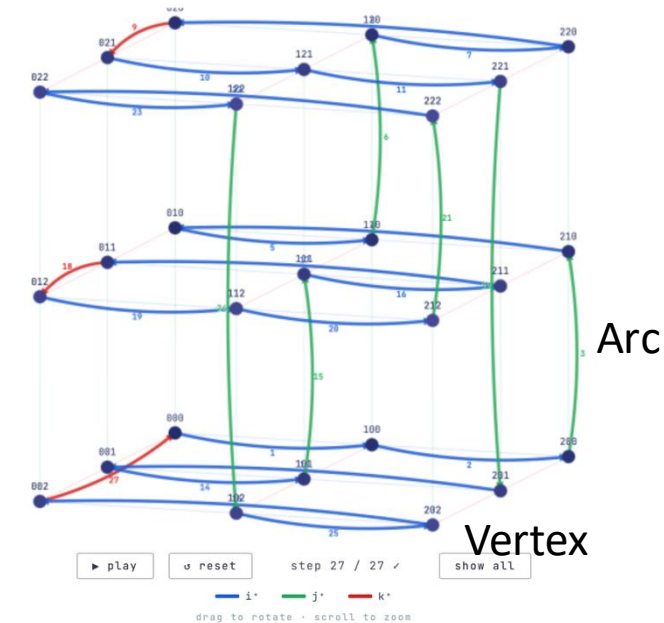
- **Exploration 1:** Claude reformulated the problem as assigning a permutation of $\{0, 1, 2\}$ to each vertex to ensure functional digraphs form single Hamiltonian cycles.
- **Exploration 2:** Claude attempted a brute-force depth-first search over permutation assignments for $m=3$, but found the search space of 6^{27} too slow without effective pruning.
- **Exploration 3:** Claude performed 2D serpentine analysis, identifying that the pattern $[0^{m-1}1]^m$ provides a valid 2D decomposition for $m=3$.
- **Exploration 4:** Claude constructed a "3D serpentine pattern" $[0^{m-1}1]^{m-1}[0^{m-1}2]^m$, but found the residual 2-regular digraph had a rigid structure that made decomposition difficult.
- **Exploration 5:** Claude analyzed the residual digraph from the previous step and identified some cycles, though they ultimately did not provide a solution.
- **Exploration 15:** Claude introduced a "fiber decomposition" that mapped arcs between layers based on $i+j+k \pmod m$.
- **Exploration 18:** Claude implemented the full fiber framework in code, finding a valid decomposition for $m=3$ in 0.1 seconds using exhaustive backtracking.
- **Exploration 20:** Claude utilized simulated annealing to find a "uniform" fiber solution for $m=4$.
- **Exploration 21:** Claude analyzed the properties of the 2D serpentine function $\sum_{i,j} \alpha Q(i,j)$ and tested it in combination with uniform choices at fibers.
- **Exploration 25:** After scaling experiments with simulated annealing failed to yield a general construction, Claude concluded that a pure mathematical approach was required.
- **Exploration 27:** Claude attempted to apply cyclic coordinate rotation to the 3D serpentine cycle, but the resulting conflicts on the hyperplane $i+j+k=m-1 \pmod m$ could not be resolved.
- **Exploration 29:** Claude proved that the "single-hyperplane + rotation" approach was impossible, determining that the direction function must use different values across a rotation orbit.
- **Exploration 30:** Claude returned to the solution found by simulated annealing in exploration 20 and observed that the fiber choice depended on only a single coordinate.



Directed Hamiltonian cycle

Directed m^3 -cycle

$m = 3$ · 27 vertices · arc pattern: $(i^+ i^+ j^+ i^+ i^+ j^+ i^+ i^+ k^+) \times 3$



Computer Science: The Knuth Problem

(28 Feb. 2026)

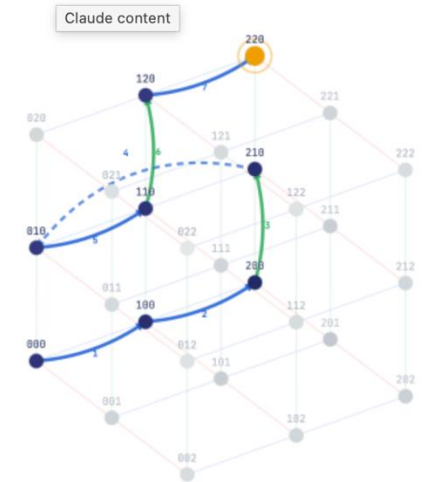
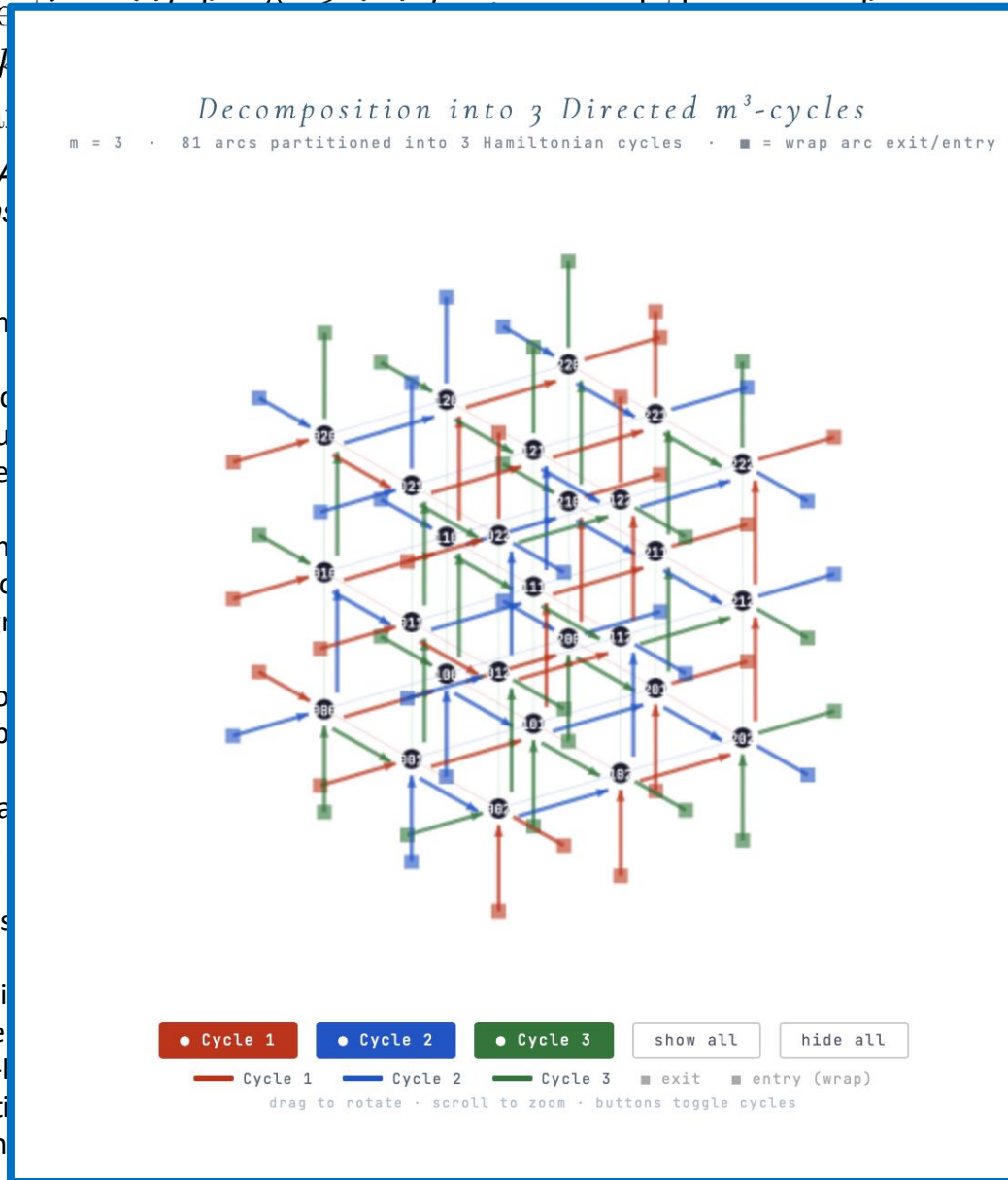
Directed m^3 -cycle

$m = 3$ · 27 vertices · all arcs go $+1 \pmod 3$ · dashed = modular wrap

Consider the digraph with m^3 vertices, one for each vertex, namely to $i+jk, ij+k$ a general decomposition of the arc

**** After EVERY exploreXX.py run, IMMEDIATELY before doing anything else. ** No exceptions until the previous one is documented here.**

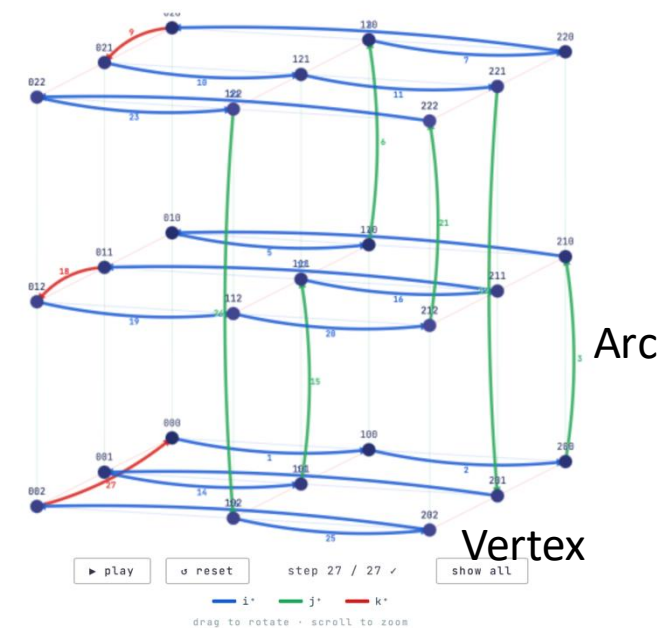
- **Exploration 1:** Claude reformulated the problem in terms of finding three Hamiltonian cycles in the digraph.
- **Exploration 2:** Claude attempted a brute-force search of the space of 6^{27} too slow without effective pruning.
- **Exploration 3:** Claude performed 2D serpentine decomposition for $m=3$.
- **Exploration 4:** Claude constructed a "3D serpentine digraph" but found it had a rigid structure that made decomposition difficult.
- **Exploration 5:** Claude analyzed the residual digraph but it did not provide a solution.
- **Exploration 15:** Claude introduced a "fiber decomposition" but it was not successful.
- **Exploration 18:** Claude implemented the full fiber decomposition using exhaustive backtracking.
- **Exploration 20:** Claude utilized simulated annealing but it did not find a solution.
- **Exploration 21:** Claude analyzed the properties of uniform choices at fibers but it did not work.
- **Exploration 25:** After scaling experiments with larger m , a more pure mathematical approach was required.
- **Exploration 27:** Claude attempted to apply cyclotomic theory but the hyperplane $\{i+j+k=m-1 \pmod m\}$ could not be used.
- **Exploration 29:** Claude proved that the "single-fiber" function must use different values across a rotation.
- **Exploration 30:** Claude returned to the solution but it depended on only a single coordinate.



Directed Hamiltonian cycle

Directed m^3 -cycle

$m = 3$ · 27 vertices · arc pattern: $(i^+ i^- j^+ i^- j^- i^+ i^- k^+) \times 3$



phs
r
d
h
a
e
e

Math: Aletheia Multi-Agent

(6 Mar. 2026)

Aletheia

A DeepMind math research agent.

Feng, T., Trinh, T.H., Bingham, G., et al. Towards Autonomous Mathematics Research. Google DeepMind, arXiv:2602.10177v2, 2026

Gemini Deep Think

Most Important Ideas and Results:

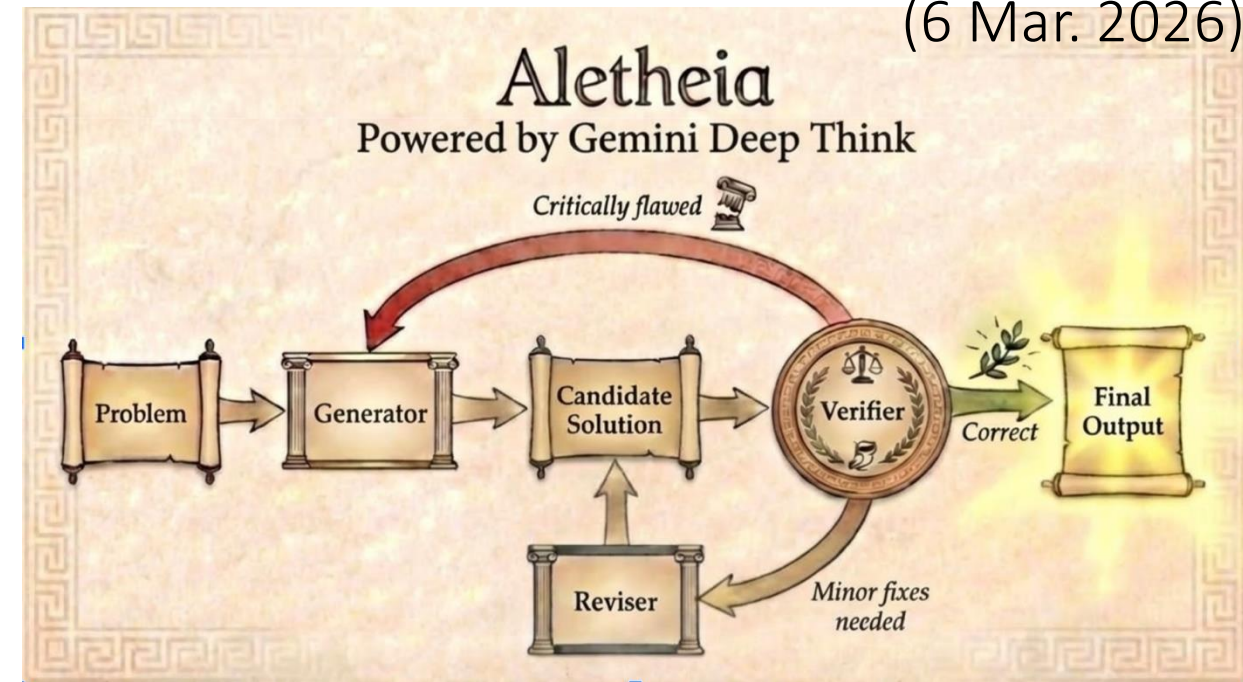
Achieve **95.1% accuracy** on the advanced IMO-ProofBench benchmark.

Agentic architecture matters: *Separating generation from verification dramatically improves reliability*—Aletheia's conditional accuracy on answered IMO problems reached 98.3%, and its ability to *admit failure* proved essential for productive human-AI collaboration.

Inference-time scaling transfers but plateaus: *Scaling compute yields substantial gains on both Olympiad and PhD-level problems, but alone is insufficient for research-grade mathematics*, motivating the agentic approach.

Tool use reduces but does not eliminate hallucinations: Internet search largely eliminated fabricated citations but shifted errors to subtler misrepresentations of real references.

Autonomous research is possible but rare and elementary: on the 700 open Erdős problems (200 solutions evaluated), only 6.5% of AI solutions were correct, and *successful cases tend to involve clever technical manipulations rather than deep creativity*.



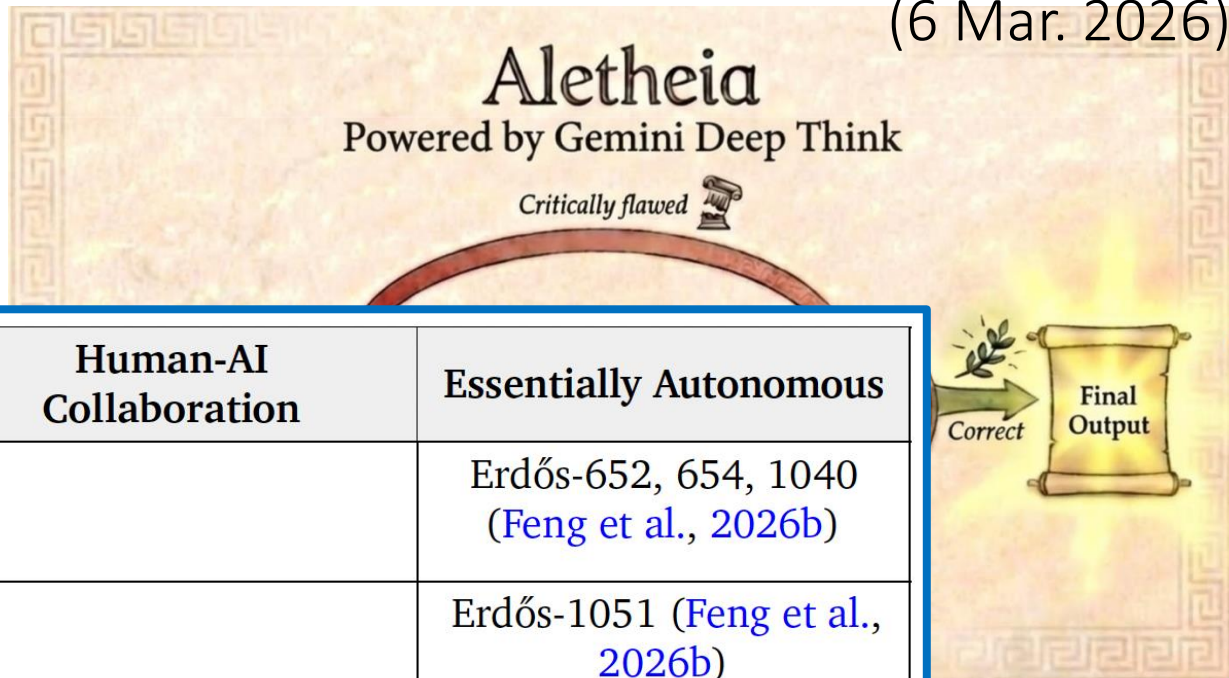
Math: Aletheia Multi-Agent

(6 Mar. 2026)

Aletheia

A DeepMind math research agent.

Feng, T., Trinh, T.H., Bingham, G.,



Gemini D

Most Im

Achieve

Agentic a

condition

productiv

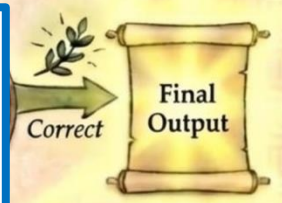
Inference

problems

Tool use

shifted e

	Primarily Human (secondary AI input)	Human-AI Collaboration	Essentially Autonomous
Level 0: Negligible Novelty			Erdős-652, 654, 1040 (Feng et al., 2026b)
Level 1: Minor Novelty			Erdős-1051 (Feng et al., 2026b)
Level 2: Publishable Research*	Complexity Bounds (ACGKMP26) Arithmetic Volumes (FYZ26)	Generalized Erdős-1051 (BKZZ26) Independence Polynomials (LeeSeo26)	Eigenweights (Feng26)
Level 3: Major Advance			
Level 4: Landmark Breakthrough			



ia's al for

PhD-level

but

Autonomous research is possible but rare and elementary: on the 700 open Erdős problems (200 solutions evaluated), only 6.5% of AI solutions were correct, and **successful cases tend to involve clever technical manipulations rather than deep creativity.**

Biomed: Google co-Scientist

(18 Feb. 2025)

Toward and AI co-scientist, arXiv:2502.18864

AI co-scientist relies on “self-play” strategies to **continuously** generates, reviews, debates, explains its reasoning and improves research hypotheses toward the research goal.

Multi-agent architecture (Agentic Models):

- All agents built from **Gemini 2.0**
- **Agents: Generation, Reflection** (peer reviewer), **Evolution, Proximity** (hypotheses), **Meta-review** (high level analysis), **Ranking** (novelty, correctness, and testability)
- Asynchronous task execution framework for **flexible compute scaling**
- **Tournament evolution process** for self-improving hypotheses generation. Feedback from the tournament creates a self-improving loop towards novel quality outputs.
- **Tools:** web search and specialized AI models to improve grounding and quality of generated research hypotheses.

Automated evaluations

Scientist-in-the-loop

Scientist

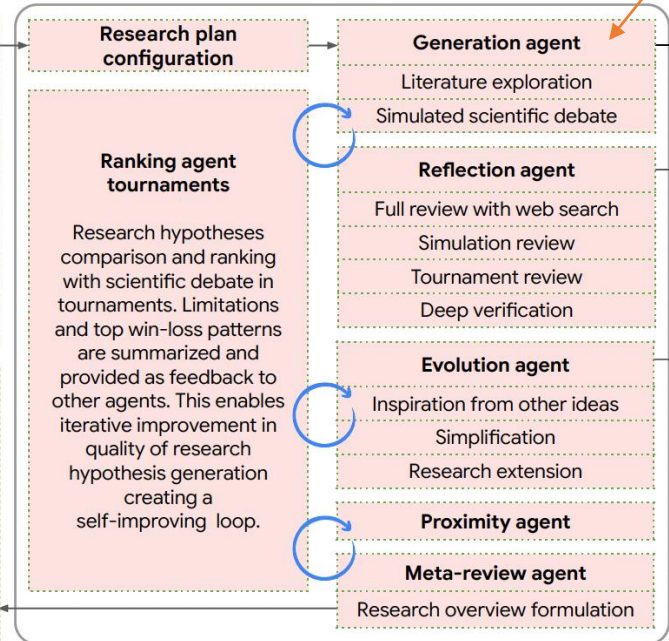
The scientist interacts with the system by specifying a research goal in natural language. They can also suggest their own ideas and proposals, provide feedback and reviews, and interact via a chat interface to guide the co-scientist system.

Discuss via chat interface

Scientist inputs



The AI co-scientist multi-agent system



Hypothesis generation

AI

AI co-scientist

The AI co-scientist continuously generates, reviews, debates, and improves research hypotheses and proposals toward the research goal provided by the scientist.

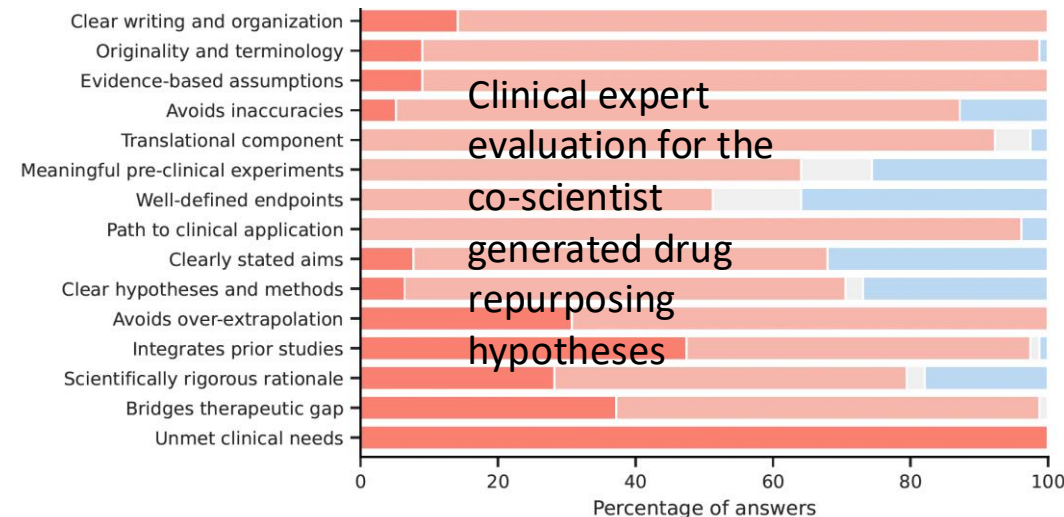
Tool Use

Search

Additional tools

Memory

3 biomed areas: **drug repurposing, novel target discovery, and explaining mechanisms of bacterial evolution and anti-microbial resistance.** Co-scientist's hypotheses for these three settings are externally, independently validated by in vitro laboratory experiments



Hypothesis Generation Tools Potential

Scientific Domains & Use Cases

Materials Science

Chemistry & Catalysis

Quantum & Advanced Physics

Earth & Subsurface Science

Fusion/Plasma Science

Accelerator Science

Nanoscience

Computing & AI

AI & Foundation Models

Robotics & Autonomous Systems

Computer Science

Applied Mathematics

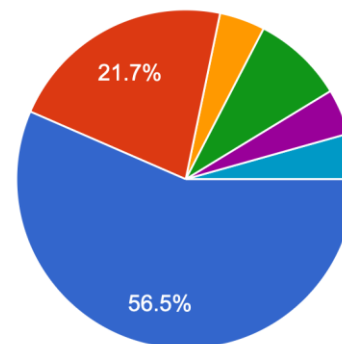
What academic degree level corresponds the most to the current level of performance of the system in your specific discipline?

23 responses

Ph. D. level in all these domains?



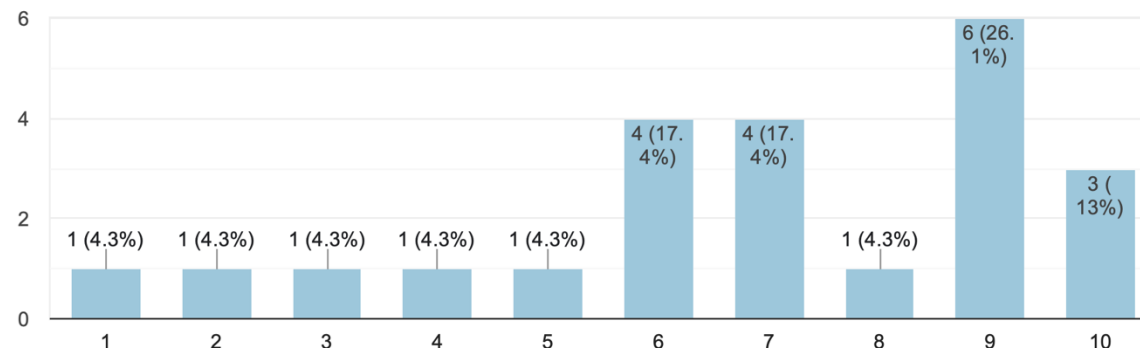
Opportunity for Many-disciplinary solutions



- Ph.D.
- Master's
- High-school
- Bachelor
- Untrustworthy PhD
- Hard to say. While the material it summarized was BS/MS level, it's inability to generate well explained ideas would have gotten a failing grade in High school.

On a scale of 1 (Not at all) to 10 (Absolutely), how likely are you to use the AI Co-Scientist if access was provided?

23 responses



State of the art

Kulkarni, A., et al. Scientific Hypothesis Generation and Validation: Methods, Datasets, and Future Directions, arXiv preprint 2505.04651v1, 2025.

The survey covers the full lifecycle: from data input and hypothesis creation through iterative refinement and real-world validation.

hypothesis generation approaches

1. knowledge-driven methods (knowledge graphs, ontologies), 2. data-driven integration (multi-omics), 3. AI-driven exploration (RAG, reinforcement learning), 4. text/concept mining, 5. simulation and modeling, 6. interactive human-AI systems, 7. causal inference, 8. dynamic knowledge systems, and 9. multi-agent systems.

hypothesis validation

- Methods: experimental testing, 2. simulation-driven validation, 3. predictive/real-time validation, 4. cross-disciplinary generalization, 5. human-AI/crowdsourced validation, 6. causal relationship validation, 7. benchmarking, 8. multi-agent validation, 9. explainability validation, 10. and hybrid methods.
- Criteria: novelty, feasibility, and hypothesis quality using mathematical formulations.
- Persistent challenges: 1. limited novelty, 2. feasibility gaps, 3. data biases, 4. lack of interpretability, 5. ethical concerns, and 6. scalability issues.

Most Important Future Directions

Enhancing Novelty: LLMs tend to reproduce established knowledge rather than generate new ideas. Generative exploration models (GANs, RL with novelty-seeking rewards), [contrastive learning](#), [cross-domain knowledge transfer](#), and [explicit novelty metrics to push beyond incremental hypotheses](#).

Improving Feasibility Assessments: Bridging the gap between computational predictions and real-world experimental outcomes is critical. Strategies include integrating [automated laboratory pipelines](#), [sim-to-real transfer techniques](#), and [feasibility-centric evaluation metrics grounded in empirical constraints](#).

Mitigating Data Biases and Ethical Risks: LLMs trained on biased corpora can produce hallucinated outputs. The paper calls for [bias detection frameworks](#), [diverse dataset curation](#), [explainable AI \(XAI\) methods](#), and [regulatory/auditing standards to ensure trustworthy, accountable AI-driven science](#).

Enhancing Interpretability and Transparency: Black-box hypothesis generation undermines trust. [Tools like IBM Watson Explainability Tools](#), [human-readable outputs](#), and [open-source model workflows](#) are essential for researcher confidence and interdisciplinary collaboration.

Encouraging High-Risk, High-Reward Hypotheses: Current frameworks penalize bold ideas. [Risk-tolerant evaluation metrics](#), ["moonshot" funding models](#), and [cross-disciplinary validation teams](#) are needed to support transformative discoveries.

Scalability and Human-AI Collaboration: [Federated learning](#), [distributed computing](#), [human-in-the-loop feedback systems](#), and [multi-agent architectures](#) are necessary to handle the growing scale and heterogeneity of scientific data while keeping domain experts meaningfully engaged in the loop.

Roadmap: A Key Element is Missing: Deep Research for Scientific Data



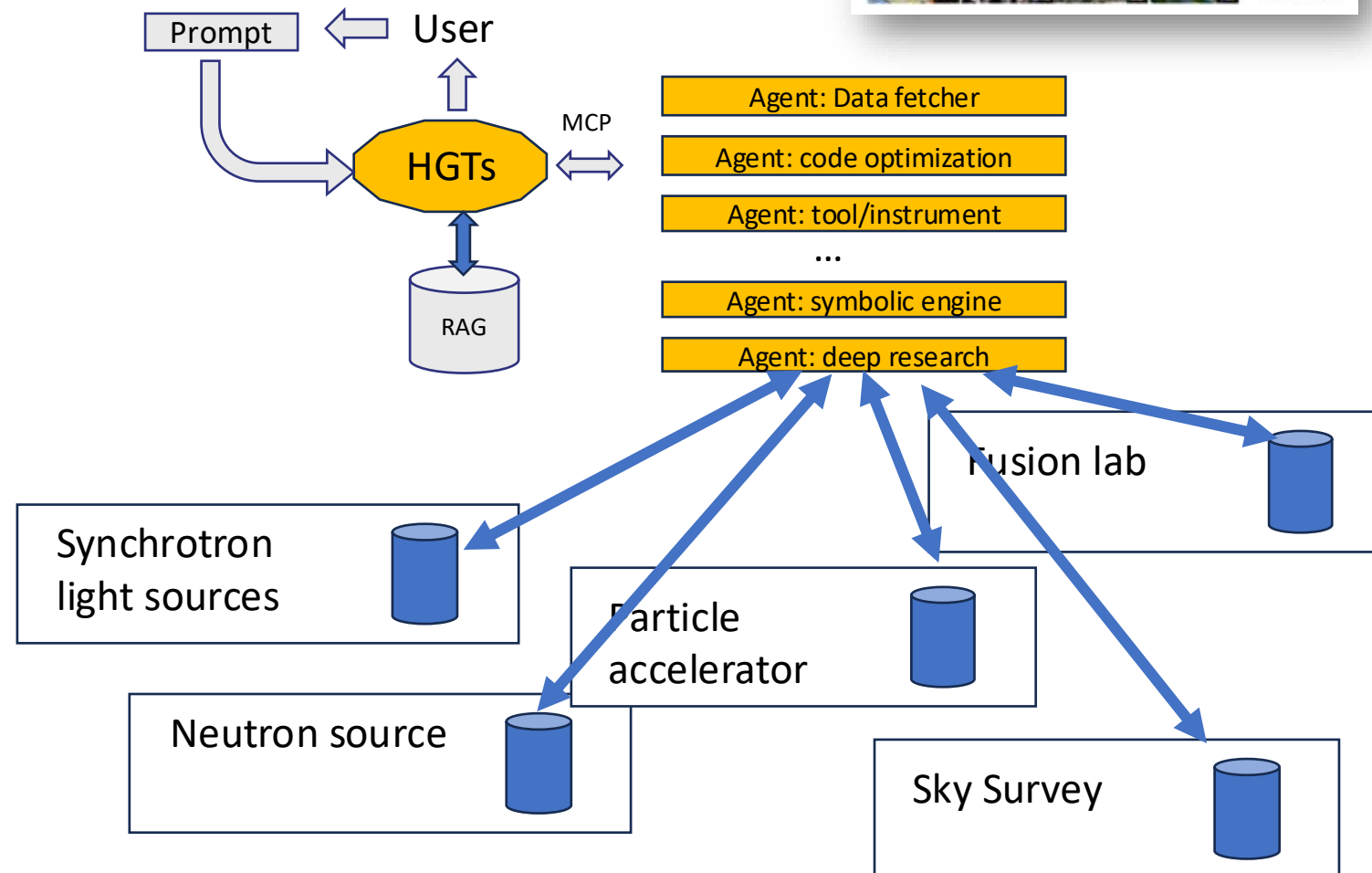
LLMs and co-scientists cannot use large scale (Tera/Peta scale) scientific data directly.

Scientific data is generated in facilities (e.g. light-sources, observatory, etc.) that are geographically dispersed.

The specific data sets needed to explore a scientific problem depends on the problem.

DeepResearch provides dynamic search and transfer of papers.

→ We need to develop the infrastructure for scientific data



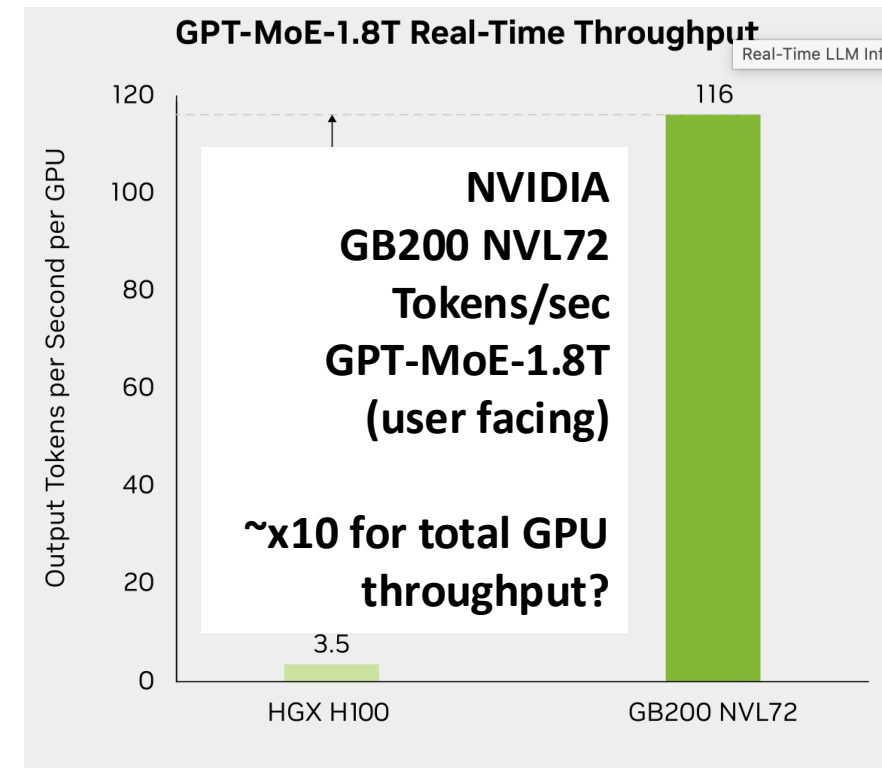
Roadmap: Hypothesis Generation may consuming a lot of tokens

Back of the envelop (need more rigorous quantification):

- For a 40,000 researchers:
 - 1 problem/month/researcher (~10/year)
 - O(1000) hypothesis per problem
 - For every hypothesis 100,000 to O(1M) tokens depending on the system complexity (LLM co-scientist), iterations limits, deep research (building a knowledge based from ~100 papers*) *3000 words/paper → 10k tokens/paper
- 40,000 x 10 x 1000 x 1,000,000 → 400 T Tokens/year (~including problem description + input data)

→ 10,000 to 100,000 GPUs (1 researcher/problem)

→ 1,000 to 10,000 GPUs (10 researchers/problem)



Roadmap: Hypothesis Generation may consume a lot of HPC resources

Back of the envelop (need more rigorous quantification):

- 40,000 Researchers (assuming 1% run large-scale simulations) → 400 Researchers
- ~10 problems per year per researcher. → 4000 problems/year
- O(1000) hypotheses / problem
- Assuming 90% hypotheses discarded by pure AI analysis (tokens) → O(100) hypotheses
- Run surrogates / digital twins to discard 90% hypotheses → O (10) hypotheses
- Run actual 0.1 Exascale simulations to discard 90% hypotheses → 1 hypothesis
- Run 1 Exascale simulations.

1 researcher / problem

- 4,000 problems/year (4m hypotheses)
- 400,000 surrogates runs / year
+ 40,000 0.1 Exaflops runs / year
+ 4000 Exaflops simulations / year

10 researchers / problem

- 400 problems/year 400,000 hypotheses
- 40,000 surrogates runs/year
+ 4,000 0.1 Exaflops runs / year
+ 400 Exaflops simulations / year

Hypothesis Generation may consume a lot of other resources

(April 23, 2026)

- $O(1000)$ hypotheses / problem
- Assuming 90% hypotheses discarded by pure AI analysis $\rightarrow O(100)$ hypotheses

Medra (San Francisco)
“physical AI scientists”:
general-purpose robot arms
with cameras mounted near
their grippers and nine
different sensors - all governed
by software that lets the arms
operate lab instruments the
way a trained human would.

<https://www.corememory.com/p/a-hundred-robots-are-running-a-bio-medra-michelle-lee>

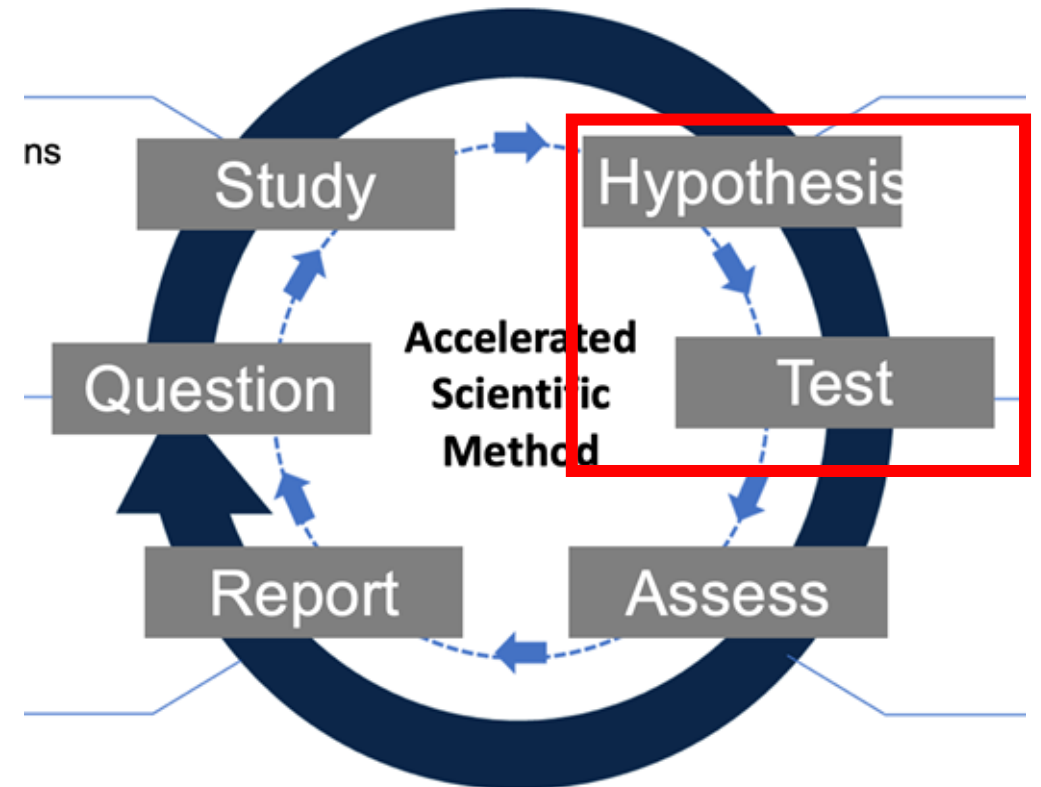


Conclusion

- Hypothesis Generation is important to consider in current road mapping efforts
- Hypotheses testing will also consume a lot of resources

Road mapping:

- Deep research for scientific data
- HGTs Impact on system workload (tokens)
- HGTs Impact on HPC workload



Thanks!

Q&As

EAIRA EVAL Methodology :
<https://arxiv.org/abs/2502.20309>

