

AI for Scientific Discovery: From Algorithmic Tools to Integrated Research Platforms

Dong-Shan Jian (Peking University)

jiandongshan@hotmail.com

AI 智能体赋能核物理研究研讨会, 中国·上海, 2026/07/04



北京大学



Outline

I. AI for established fields: current status

II. AI for frontier fields: challenges

- a. Complexity of frontier research tasks**
- b. The fundamental research–coding gap**
- c. Accessibility for the broader research community**

III. Aether: an integrated vibe research platform

IV. Summary and outlook

Primary focus: reasoning

- Primary focus: 2023 ~ 2025
- LLMs **hallucinate** in complex (long) scientific reasoning tasks
- Scientific research has extremely stringent requirements for **accuracy**
- Inherent complexity of natural science
 - Modeling of real-world scenarios
 - Understanding of fundamental laws
 - Long, multi-step and unstructured reasoning
 - Implicit constraints
 - Deterministic & probabilistic, precise & approximate
 - ...

Various approaches

➤ **Training-based: improve the model's intrinsic reasoning**

- Scaling
- Fune-tuning
- Reinforcement learning

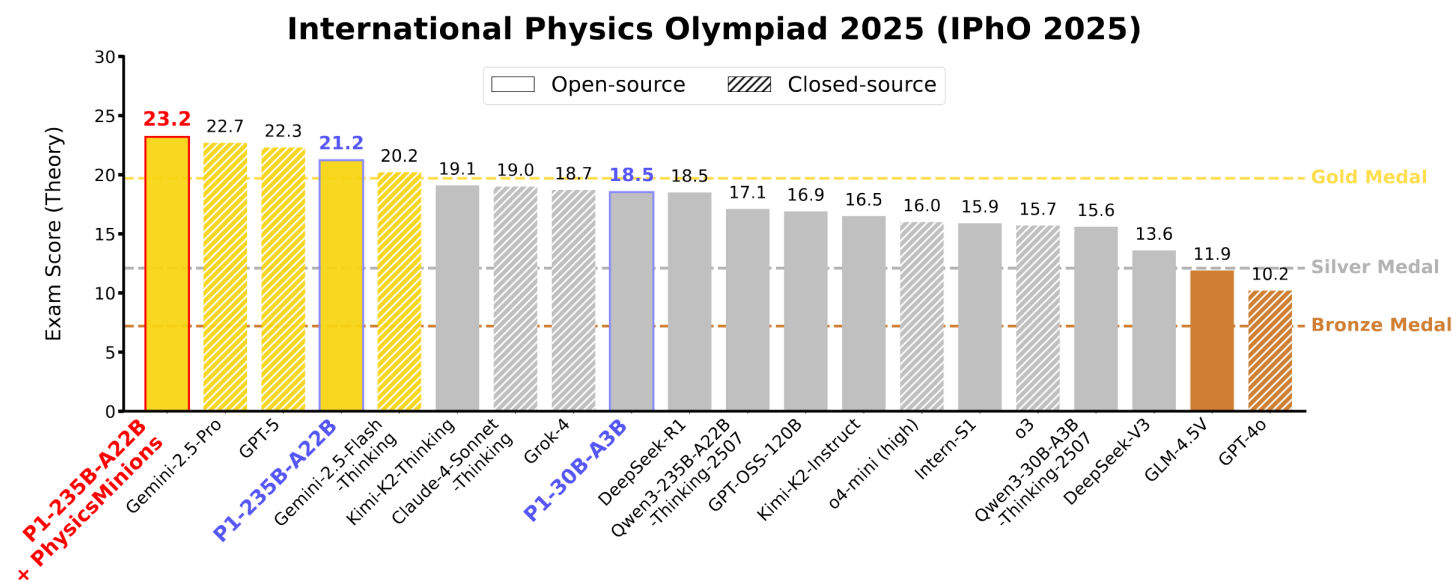
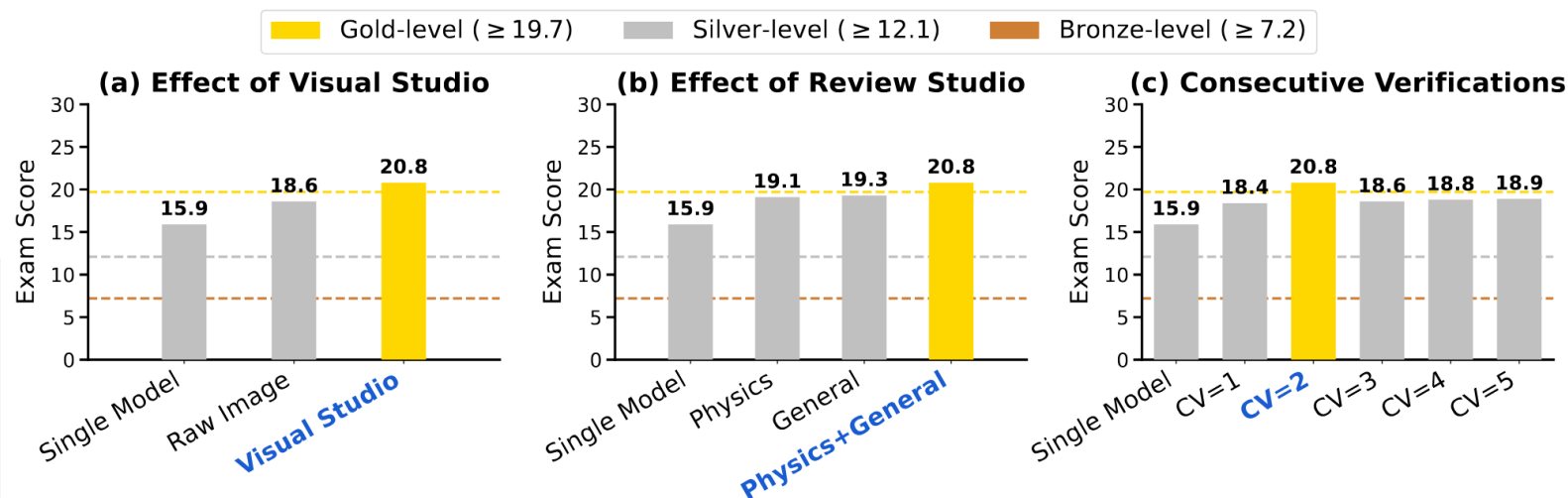
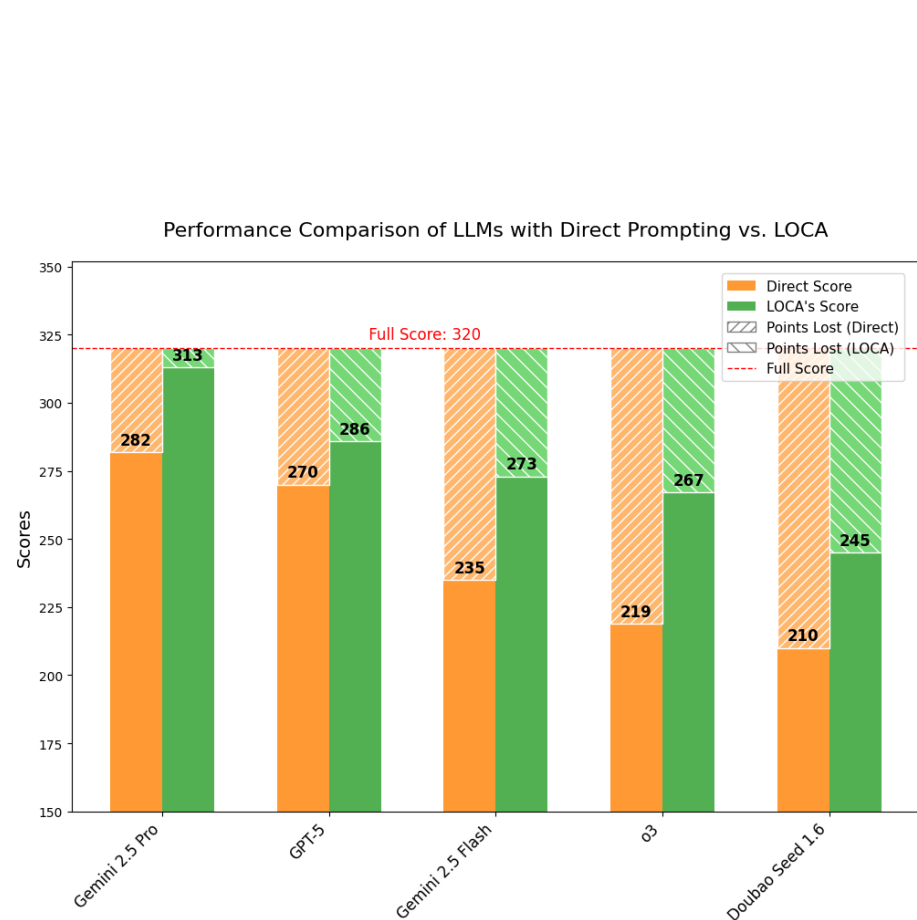
➤ **Training-free: augment reasoning with external mechanisms**

- Prompt engineering
- Tool calling
- External knowledge
- Workflows / agents (early-stage)
- Domain-specific design
- E.g., CoT, ToT, GoT, self-refine, MAD, Physics Supernova, LOCA...

➤ **Breakthrough progress, significantly surpassing humans**

Achievements

➤ E.g., on Olympiad-level physics problems



Outline

I. AI for established fields: current status

II. AI for frontier fields: challenges

a. Complexity of frontier research tasks

b. The fundamental research–coding gap

c. Accessibility for the broader research community

III. Aether: an integrated vibe research platform

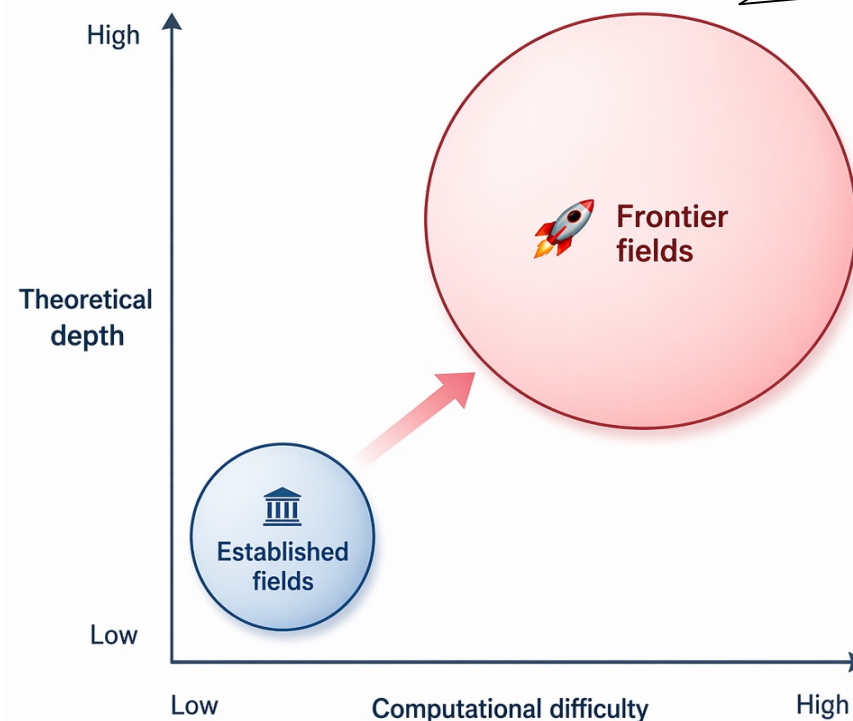
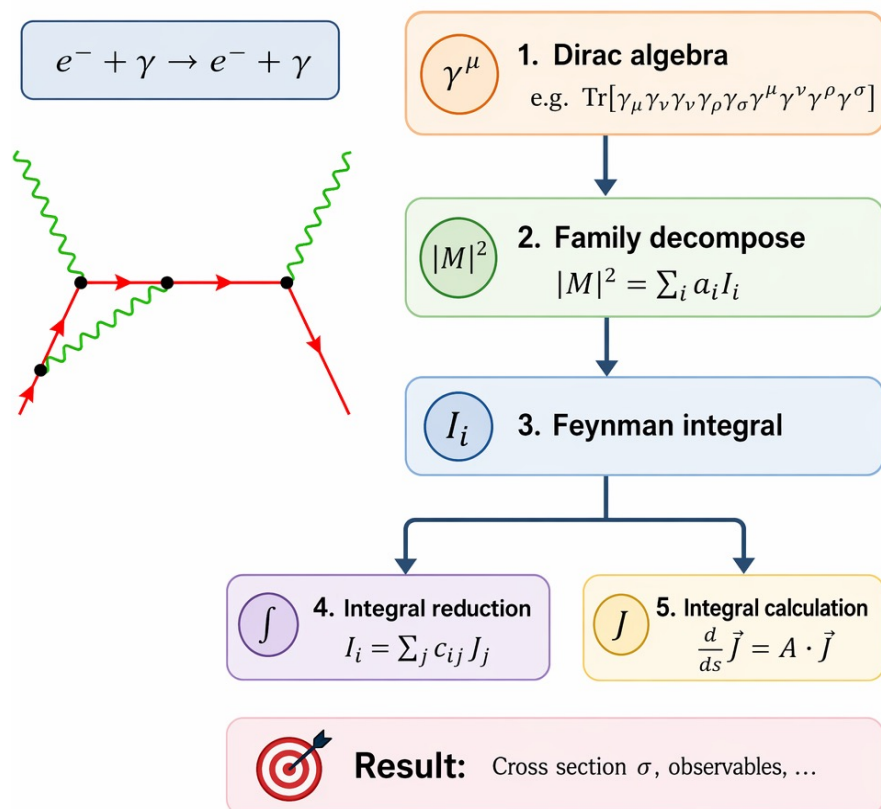
IV. Summary and outlook

Challenges for frontier fields

➤ Great uncertainty

Highly complex domain-specific methods

➤ Example: perturbative quantum field theory



Multiple
methods
integration

Solution 1: prompt engineering

➤ Guide AI's reasoning behavior

You will be instructed to convert the noninteracting Hamiltonian $\{nonint_symbol\}$ in the second quantized form from the basis in real space to the basis in momentum space. To do that, you should apply the Fourier transform to $\{real_creation_op\}$ in the real space to the $\{momentum_creation_op\}$ in the momentum space, which is defined as $\{definition_of_Fourier_Transformation\}$, where $\{real_variable\}$ is integrated over all sites in the entire real space. You should follow the EXAMPLE below to apply the Fourier transform. [Note that hopping have no position dependence now.] You should recall that $\{nonint_symbol\}$ is $\{expression_nonint\}$. Express the total noninteracting Hamiltonian $\{nonint_symbol\}$ in terms of $\{momentum_creation_op\}$. Simplify any summation index if possible.

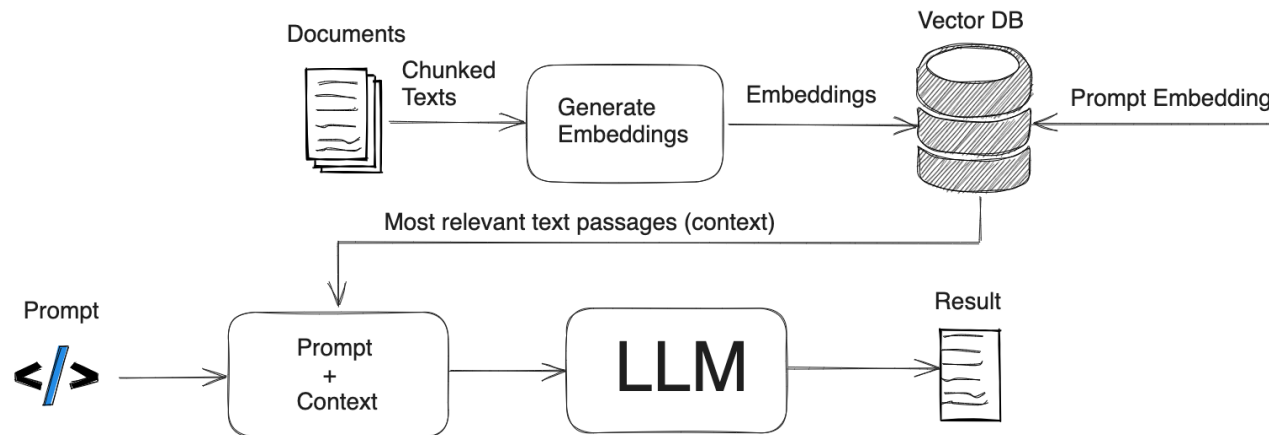
Use the following conventions for the symbols (You should also obey the conventions in all my previous prompts if you encounter undefined symbols. If you find it is never defined or has conflicts in the conventions, you should stop and let me know): $\{definition_of_variables\}$

<https://doi.org/10.1038/s42005-025-01956-y>

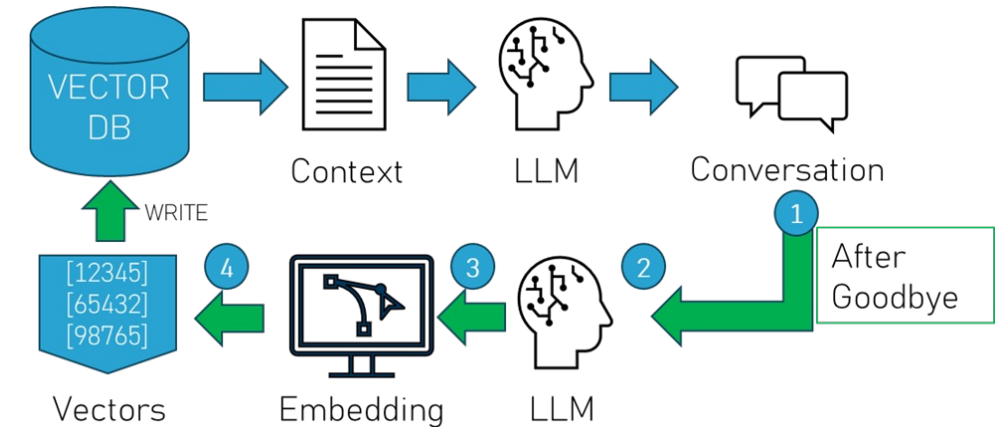
Solution 2: RAG / memory

➤ Provide AI with external knowledge that it hasn't encountered

➤ **RAG**(Retrieval-Augmented Generation)



➤ **Memory**(accumulated experience)



➤ **Difference**

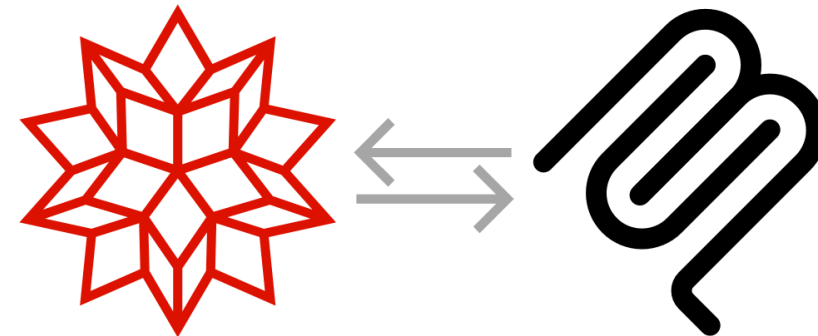
	Memory	RAG
Data updates	atomic, frequent	batch, infrequent:
Primary use	personalized, dynamically evolving facts	general-purpose, relatively static domain knowledge

Solution 3: tool calling/MCP

➤ Tool calling

Considering absolute accuracy and efficiency

Leave professional work to professional tools

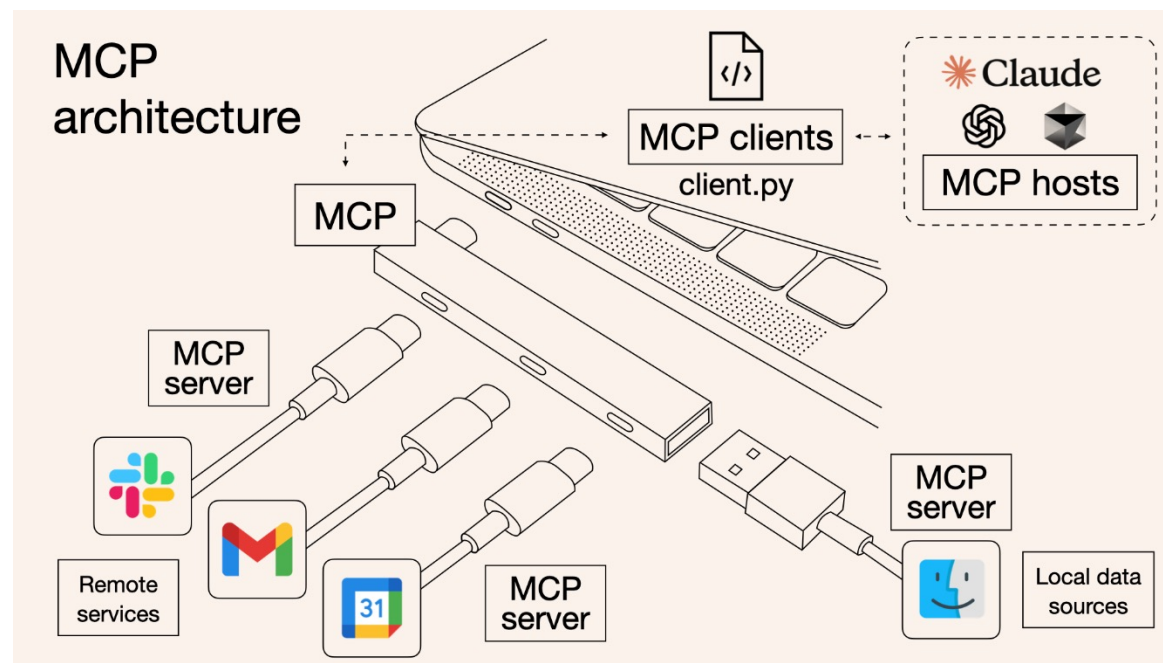


<https://resources.wolframcloud.com/PacletRepository/resources/Wolfram/MCPServer/>

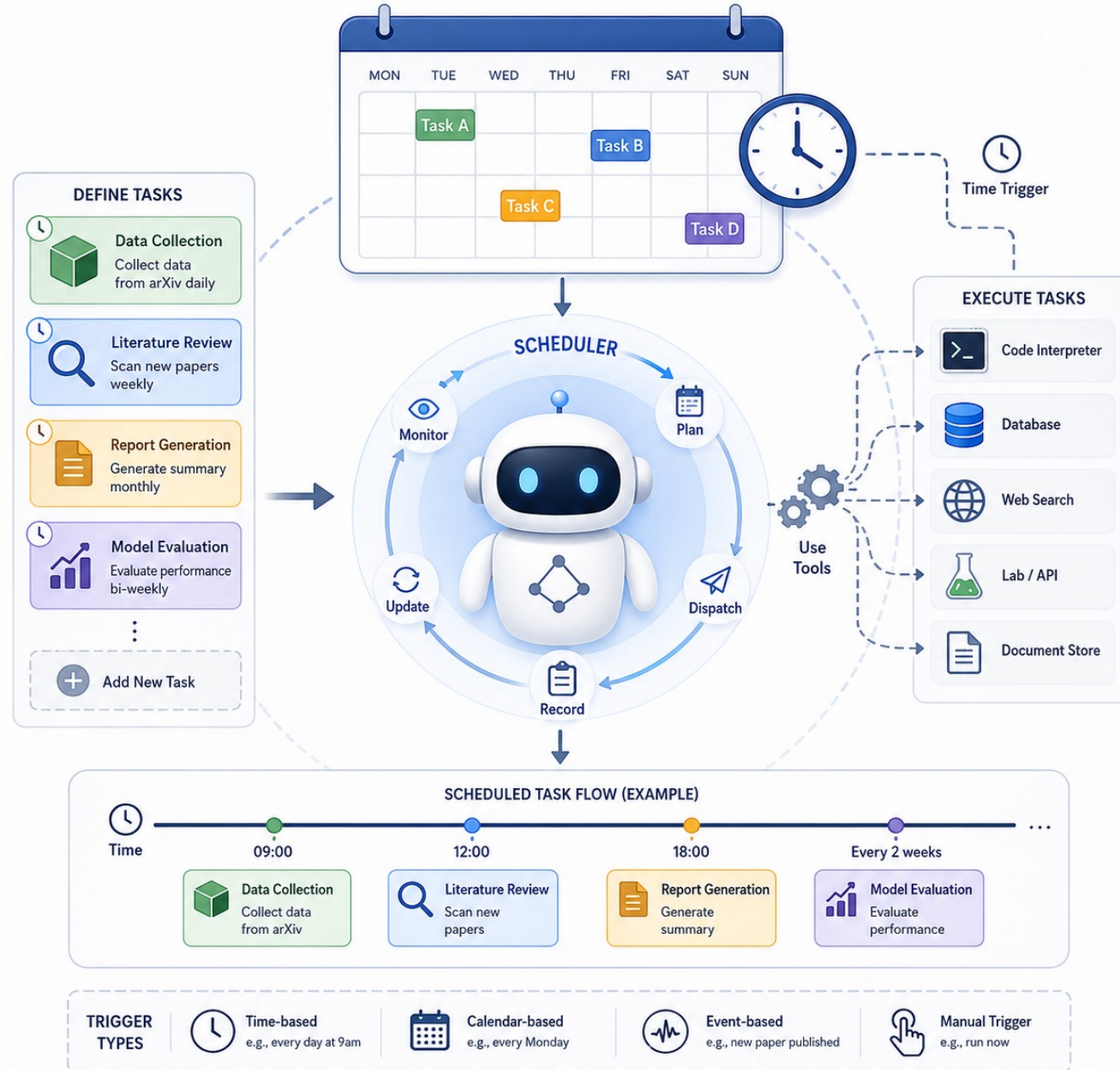
➤ MCP(Model Context Protocol)

Standard protocol for communication between LLMs and tool services

- What tools?
- How to use?
- When to call?

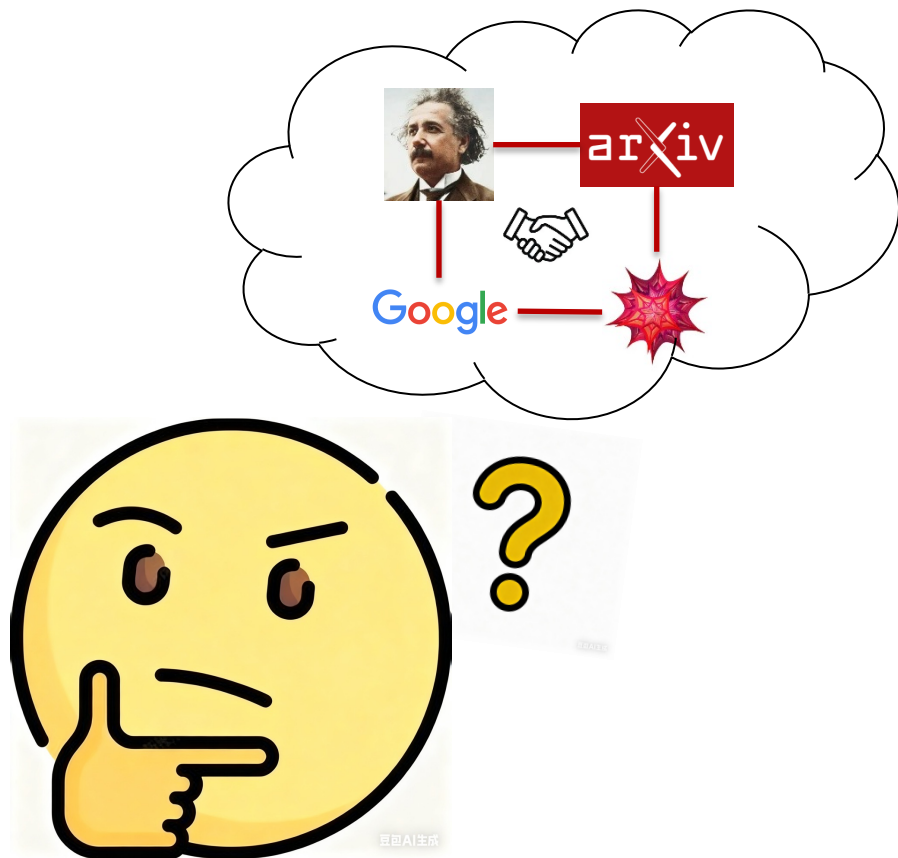


Solution 4: scheduling

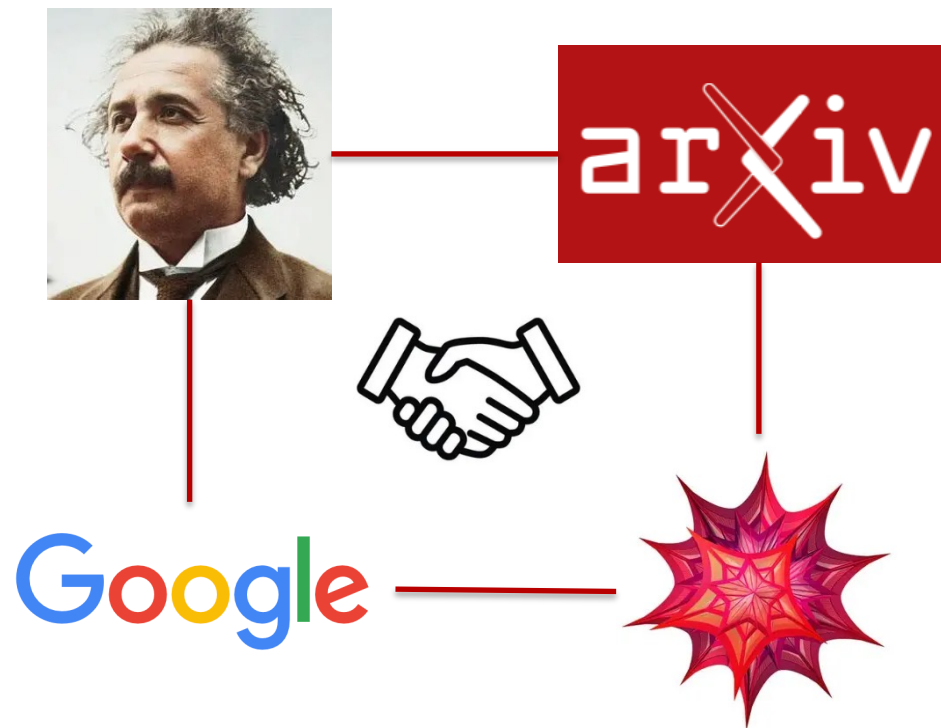


Solution 5: multi-agent workflow

➤ Single-LLM belike:



➤ Multi-agent belike:

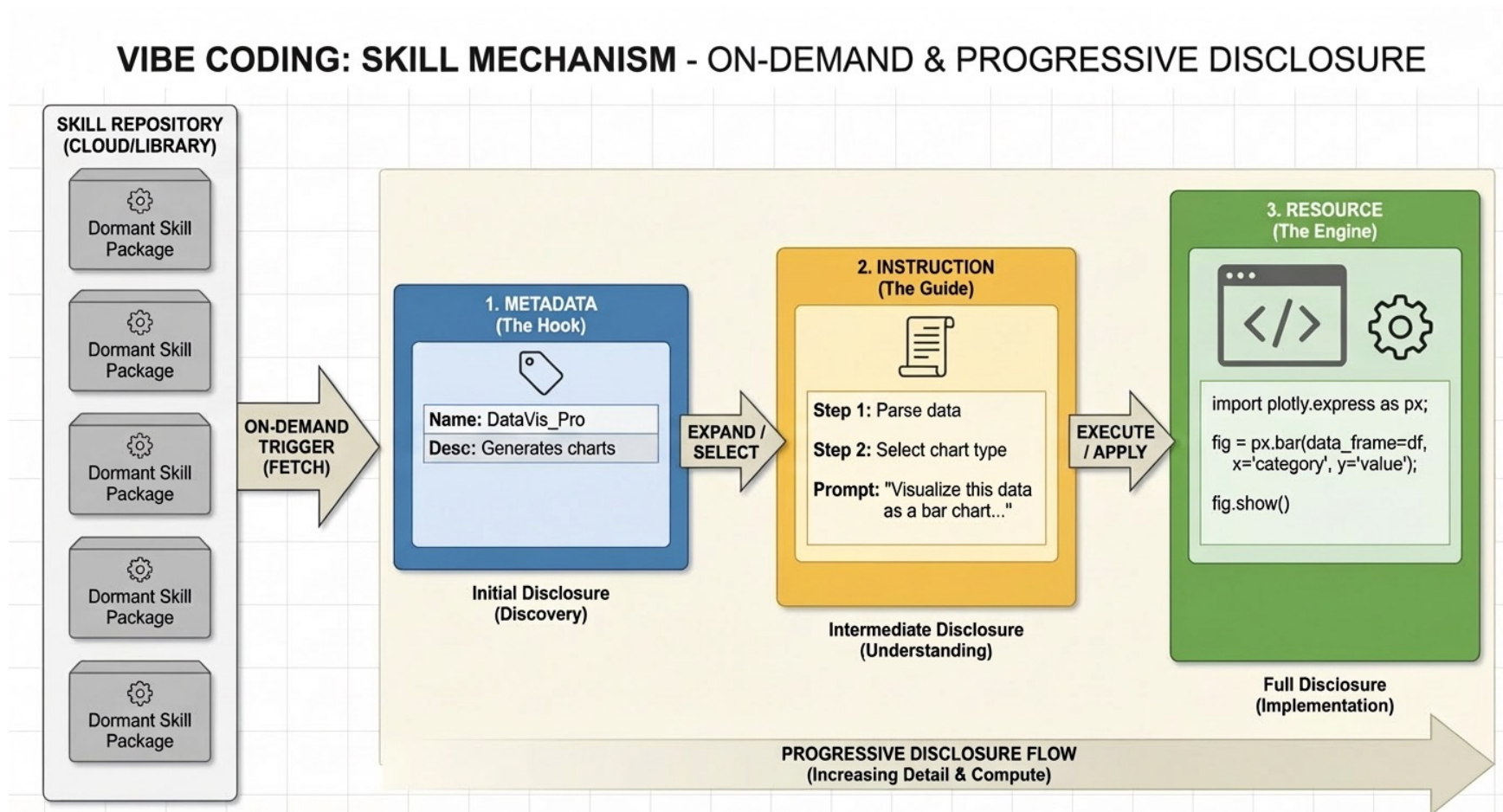


➤ For truly complex, multi-step problems

➤ Agents with different roles collaborate like humans

Solution 6: skills

- Proposed by Anthropic in 2025
- Program written in natural language



Outline

I. AI for established fields: current status

II. AI for frontier fields: challenges

a. Complexity of frontier research tasks

b. The fundamental research–coding gap

c. Accessibility for the broader research community

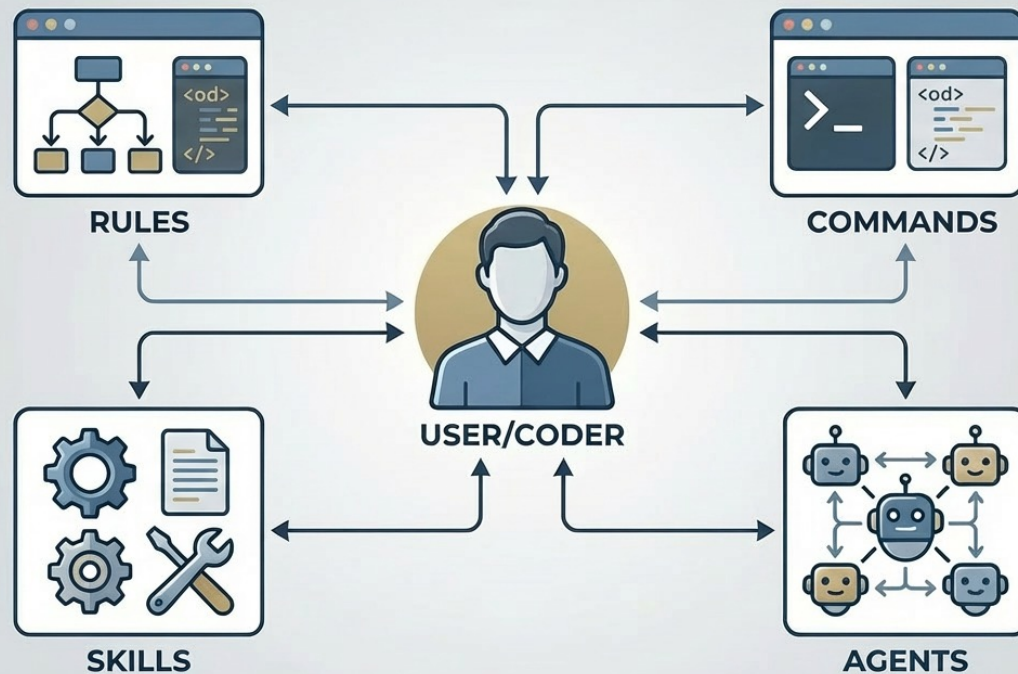
III. Aether: an integrated vibe research platform

IV. Summary and outlook

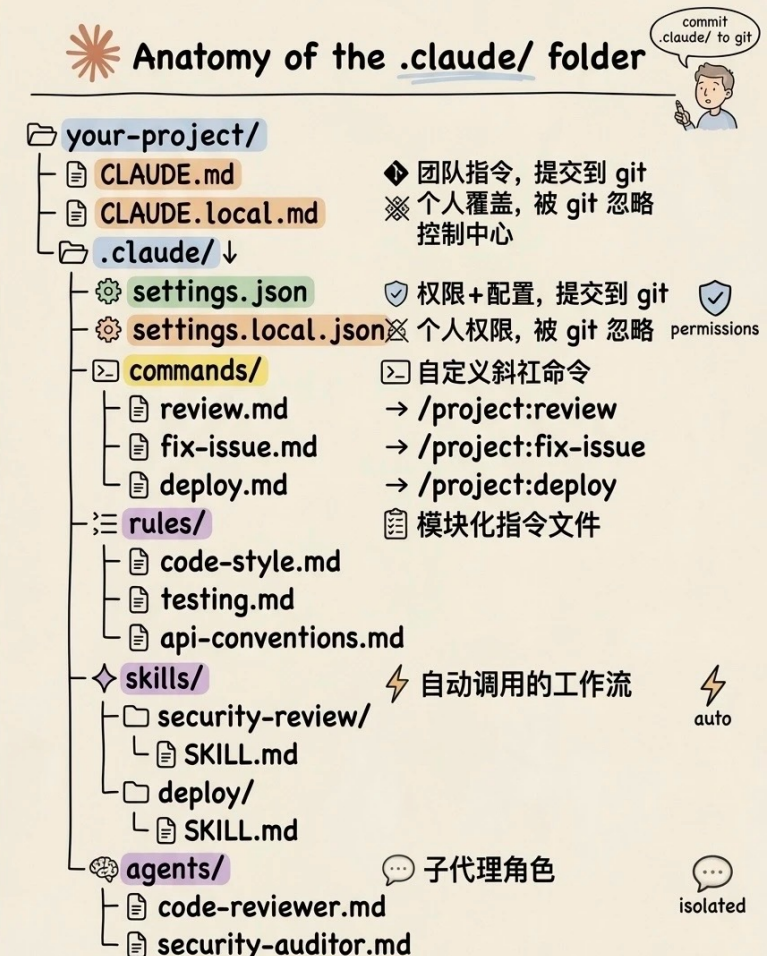
From vibe coding to vibe research

- Inspiration from vibe coding: a high level of automation
- Coding vs. research: fundamental differences in underlying logic

VIBE CODING: UNDERLYING LOGIC

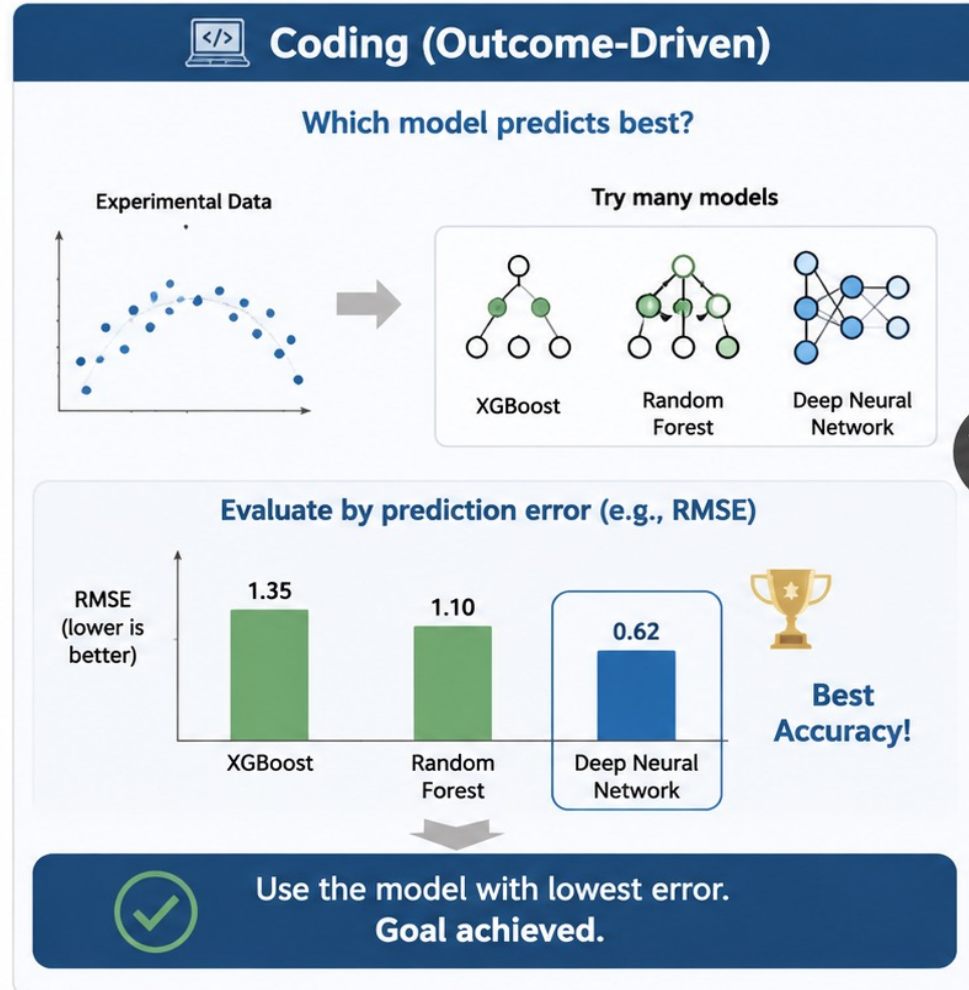


Anatomy of the .claude/ folder

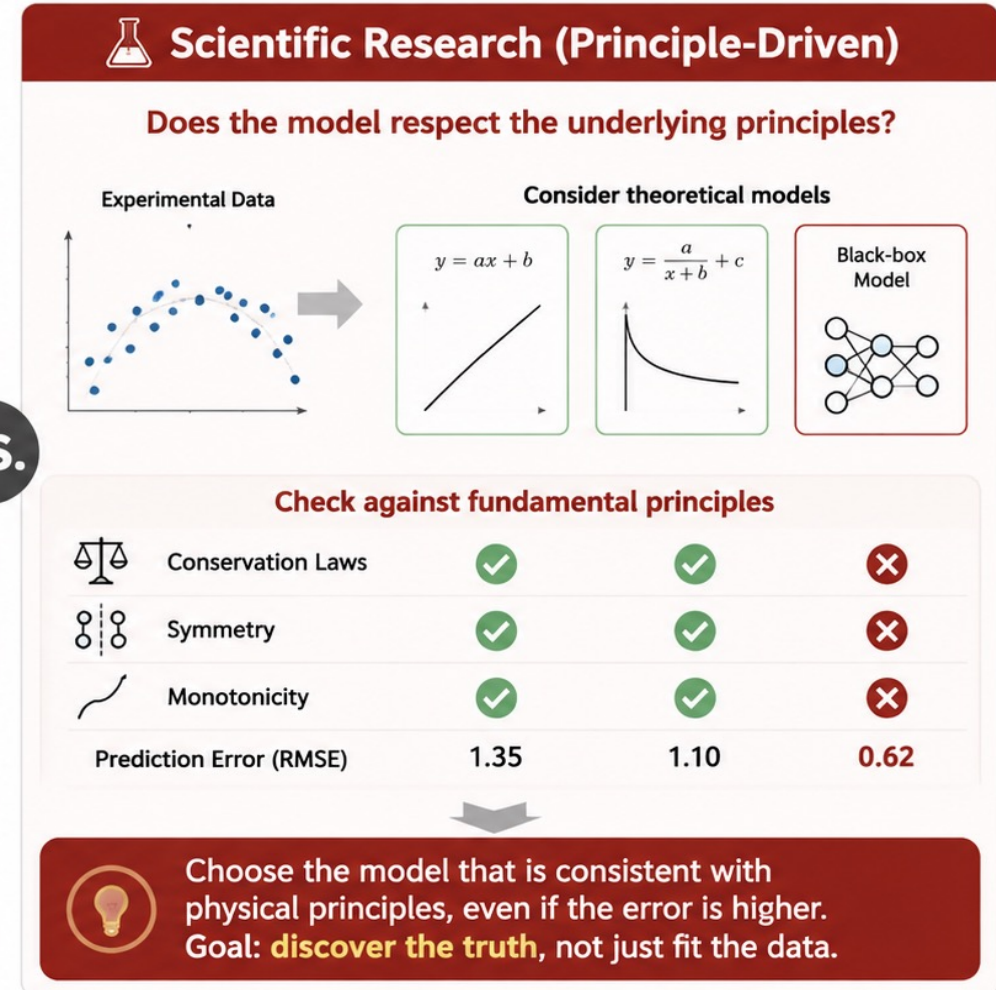


The fundamental research-coding gap

➤ Outcome-driven vs. principle-driven



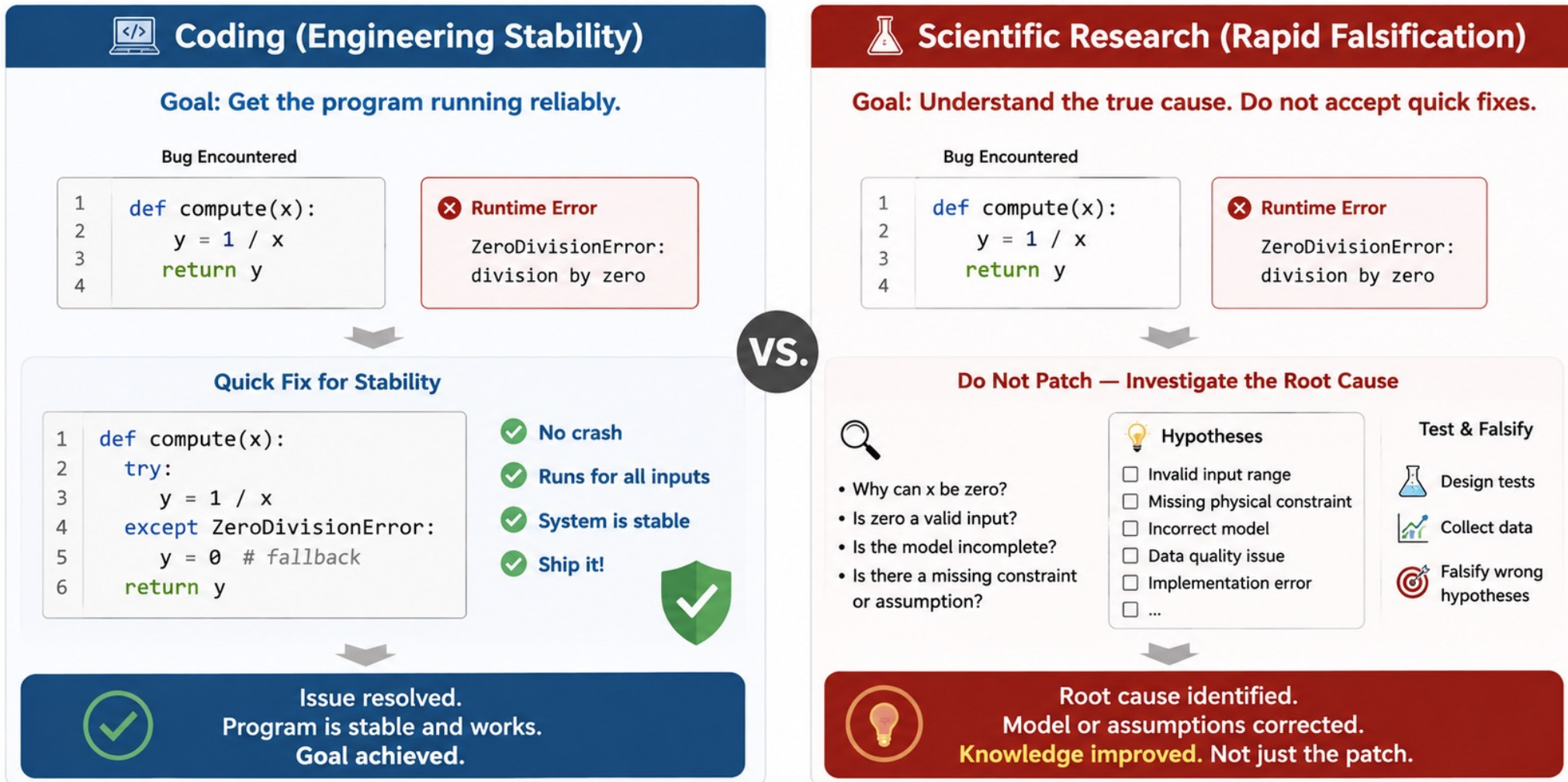
VS.



Key Difference: Engineering optimizes for the best prediction. Science seeks models that are true to reality.

The fundamental research-coding gap

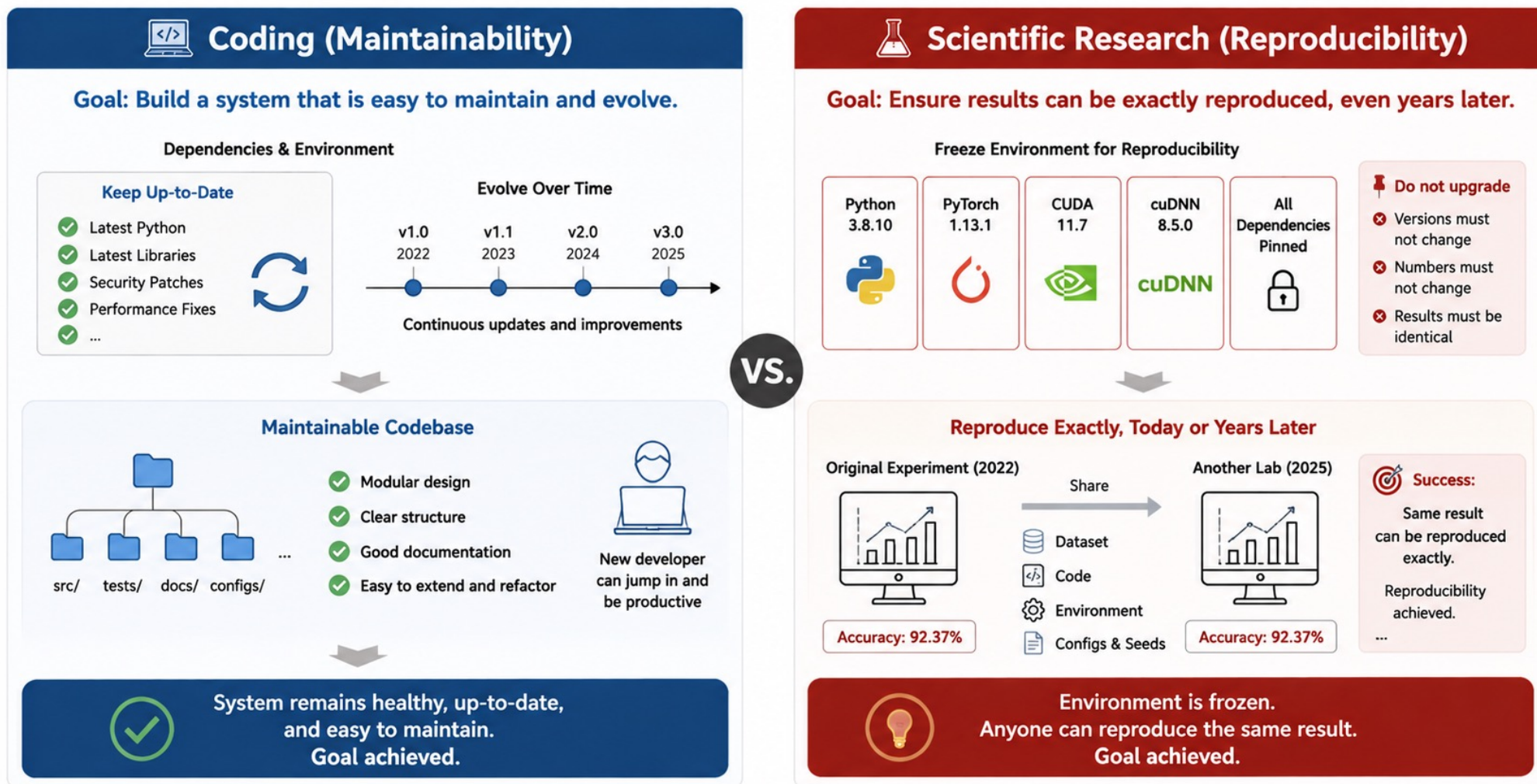
➤ Engineering stability vs. rapid falsification



Key Difference: Engineering fixes failures. Science explains failures. We falsify to learn the truth.

The fundamental research-coding gap

➤ Maintainability vs. reproducibility



Key Difference: Engineering evolves systems over time. Science preserves environments to reproduce the same results.

Outline

I. AI for established fields: current status

II. AI for frontier fields: challenges

a. Complexity of frontier research tasks

b. The fundamental research–coding gap

c. Accessibility for the broader research community

III. Aether: an integrated vibe research platform

IV. Summary and outlook

Accessibility for the broader research community

Open Source VS Closed Source



- ❌ Proprietary & opaque
- ❌ No transparency
- ❌ Limited customization
- ❌ Vendor lock-in
- ❌ Restrictive licensing

I can't see what's inside...



- ✅ Open & transparent
- ✅ Auditable & trustworthy
- ✅ Customizable & extensible
- ✅ No vendor lock-in
- ✅ Research-friendly license

I understand, I can modify, I can build on it.



Graphical Interface VS Terminal Commands



- ❌ Command-line only
- ❌ Hard to learn
- ❌ Error-prone
- ❌ Unfriendly for beginners
- ❌ High entry barrier

What does this command even mean?



```
$ git clone ...  
$ pip install ...  
$ python run.py --config  
$ screen -r session  
$ tail -f logs.txt  
$ ...
```



- ✅ Intuitive & visual
- ✅ Easy to get started
- ✅ Fewer errors
- ✅ Friendly for all researchers
- ✅ Lower entry barrier

Everything is clear and just a click away.



Flexible Model Selection VS Fixed Model Choice



- ❌ Locked to one model
- ❌ No freedom to switch
- ❌ May not fit every task
- ❌ Dependent on provider
- ❌ Limited innovation

It works sometimes, but not for my task...



Only one option



- ✅ Choose the best for the task
- ✅ Switch anytime
- ✅ Support multiple models
- ✅ More options, better results
- ✅ Encourages innovation

I can use the best model for each task.



Pick the right model

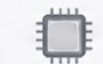


Remote Command VS Local Constraints



- ❌ Limited by local hardware
- ❌ Hard to access remotely
- ❌ No persistent environment
- ❌ Difficult collaboration
- ❌ Not always on, not always connected

My laptop is slow, I'm away from my desktop, and my runs keep failing...



Local Machine

Limited Resources

Disconnected

- ✅ Run anywhere, anytime
- ✅ Powerful remote resources
- ✅ Persistent & reliable environment
- ✅ Easy collaboration & sharing
- ✅ Always-on, always-accessible

I can run, monitor, and collaborate from anywhere.



From anywhere, on any device



Laptop

Tablet

Phone

Cloud / Remote Server



Accessible, Flexible, and Powerful — For Every Researcher.



Open and transparent.



Easy to use.



Flexible and intelligent.

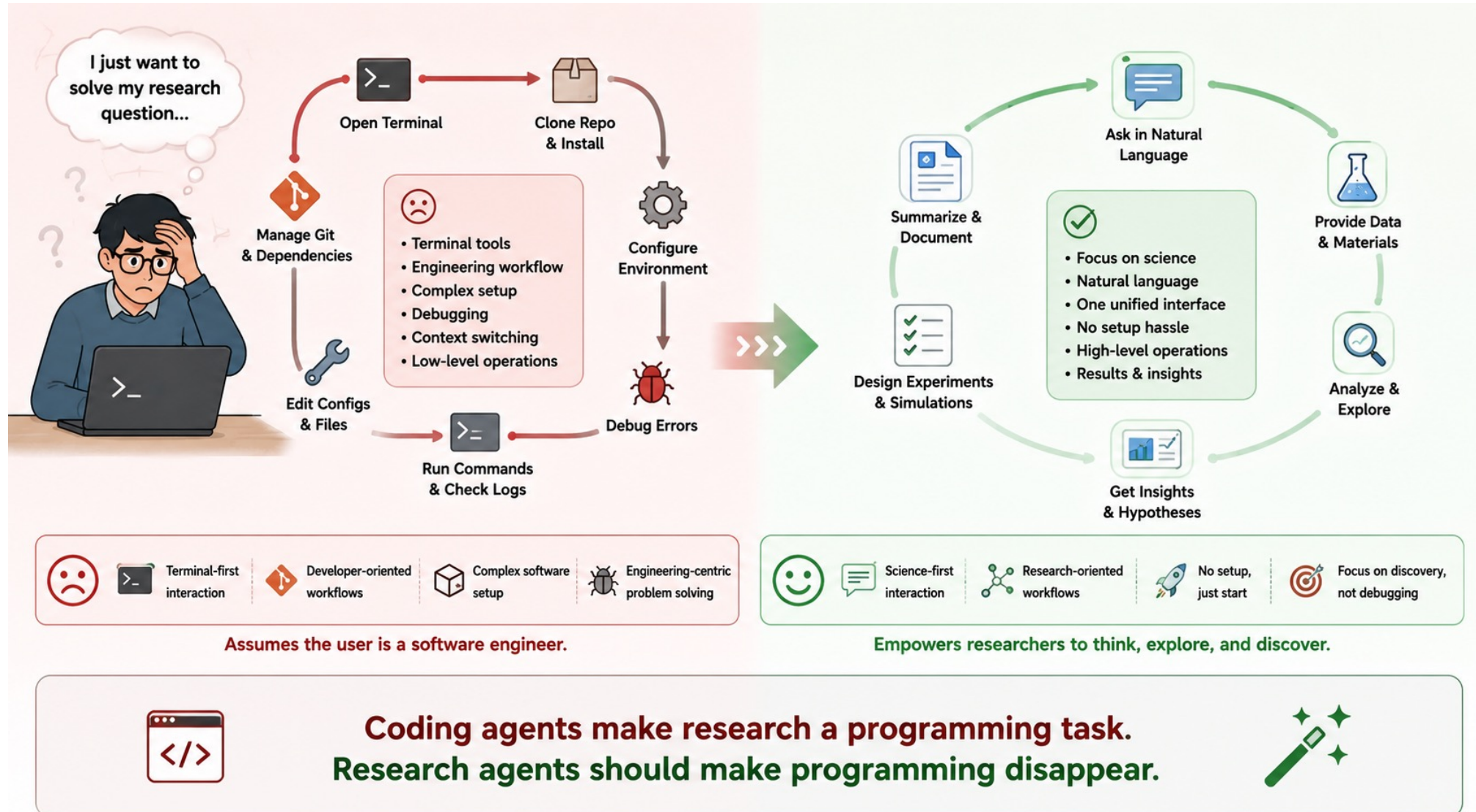


Anywhere, anytime.

AI for research should be accessible to all, not just a few.

Accessibility for the broader research community

➤ Not all researchers are skilled programmers



Outline

I. AI for established fields: current status

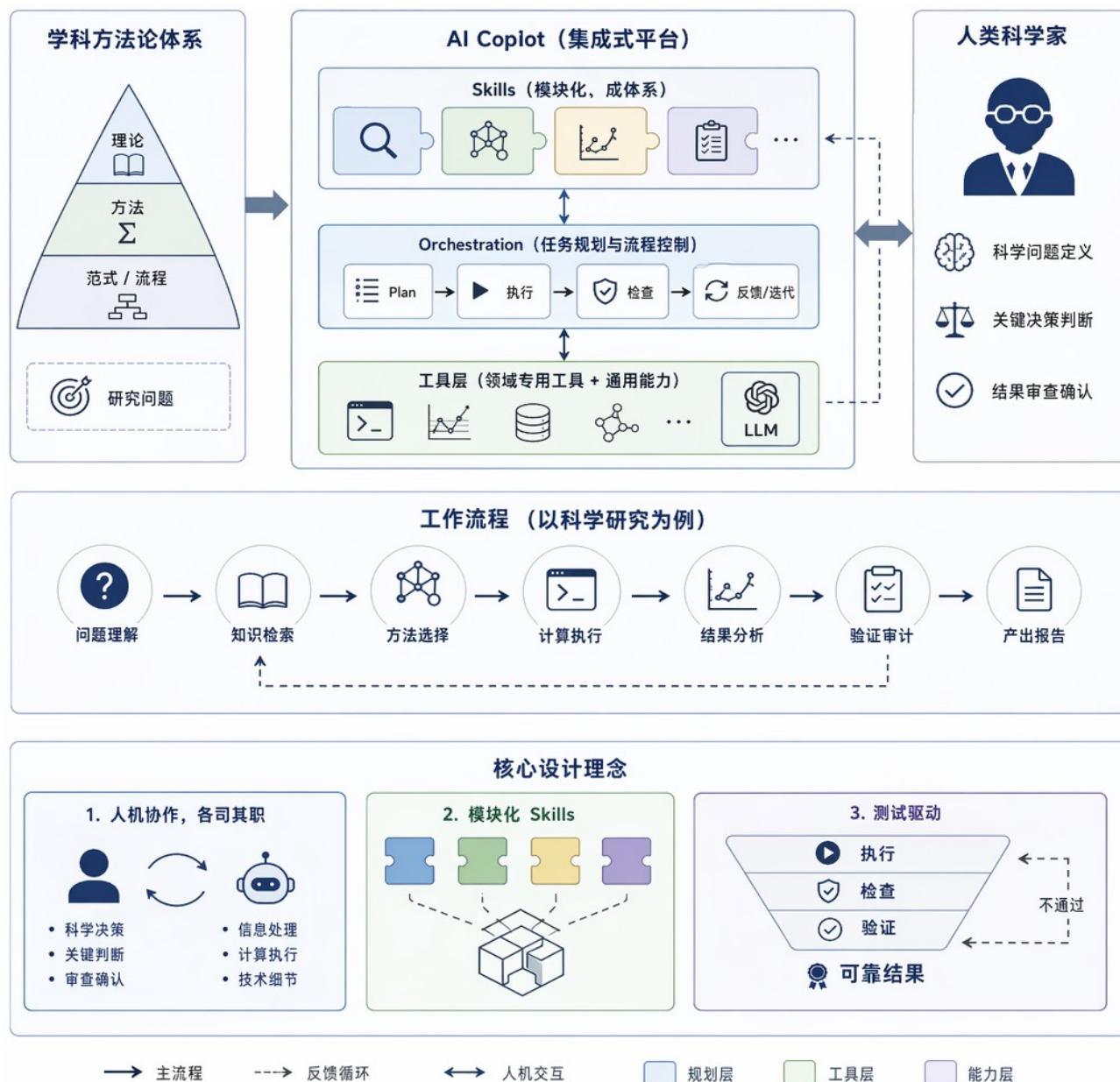
II. AI for frontier fields: challenges

- a. Complexity of frontier research tasks
- b. The fundamental research–coding gap
- c. Accessibility for the broader research community

III. Aether: an integrated vibe research platform

IV. Summary and outlook

From disciplinary methodology to AI copilot



Aether

- Autonomous Engine for Theoretical & Hands-on Exploration in Research
- To achieve a more suitable kind of vibe research
- Based on OpenCode

<https://aether.aiphys.cn/>

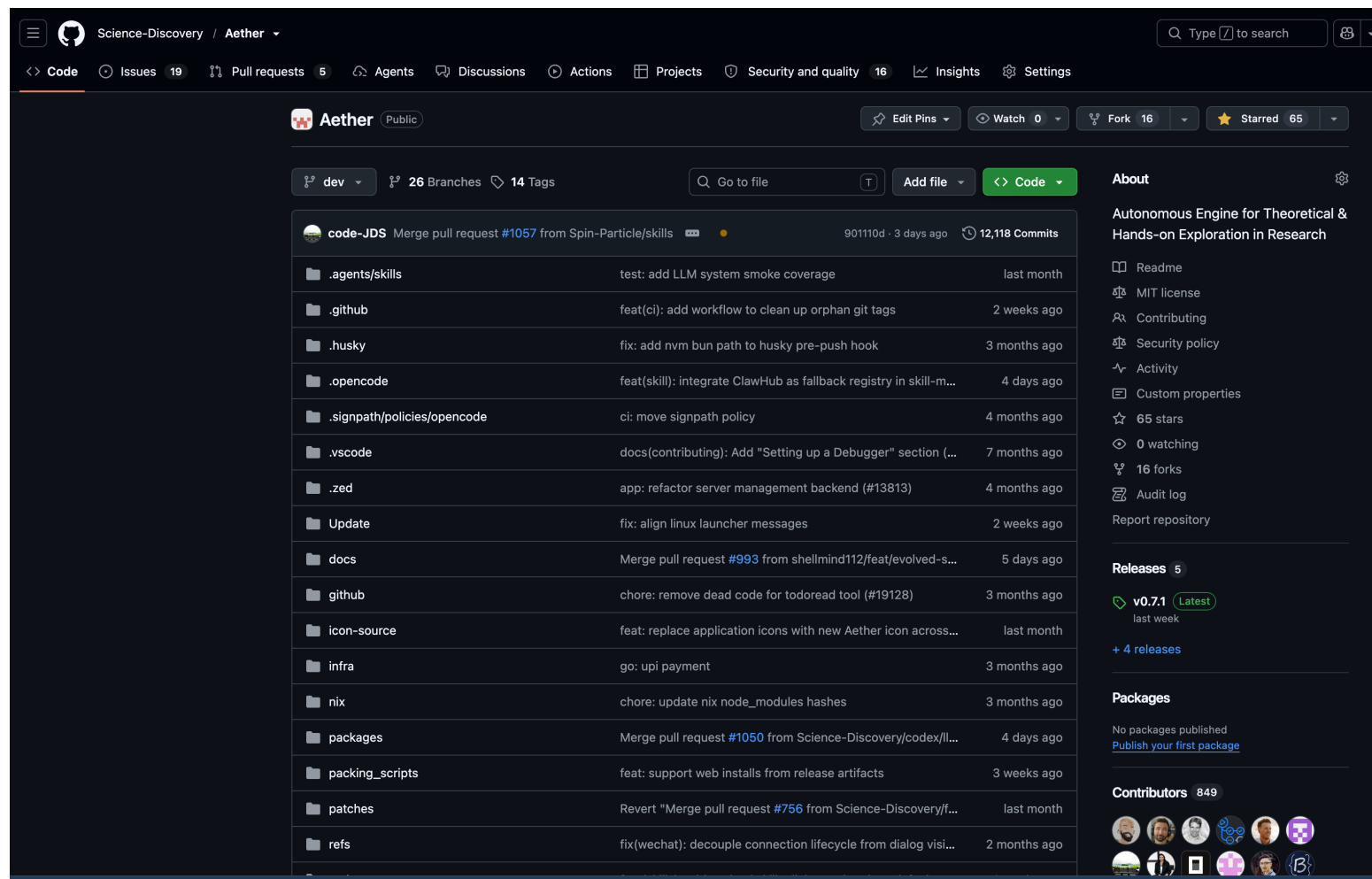
<https://github.com/Science-Discovery/Aether>

The screenshot displays the Aether website interface. The main content area features the title 'Aether: 专为科研人员定制的 AI助手' (Aether: AI Assistant Customized for Researchers) in a large, bold font. Below the title, a paragraph describes its capabilities: '强大的编程和研究能力，科研全流程护航，兼容各种模型与提供商，支持微信、QQ、飞书一键接入，开箱即用，让科研触手可及。' (Powerful programming and research capabilities, full-process research protection, compatible with various models and providers, supporting one-click access to WeChat, QQ, and Feishu, ready to use out of the box, making research accessible). Three buttons are provided: '立即查看 GitHub' (View GitHub immediately), '下载客户端' (Download client), and '查看微信连接教程' (View WeChat connection tutorial). A link to '查看文档' (View documentation) is also present. At the bottom, a contact information box states: '如有问题或建议，请联系 aethersci@163.com' (If you have any questions or suggestions, please contact aethersci@163.com).

The sidebar on the right, titled '最新动态' (Latest News), contains two entries. The first entry, dated '2026-06-04', is a '发布' (Release) announcement for version 'v0.7.0'. It highlights 'UI更新，新增skill演化、持久化memory机制与项目级数据库' (UI update, new skill evolution, persistent memory mechanism, and project-level database). The text describes three core features: 'Skill Evolution' for automatic skill management, a 'persistent memory system' for cross-session knowledge accumulation, and a 'project-level database' for improved stability and isolation. It also mentions new features like a task scheduler, MaaS model provider, and session history search. The second entry, dated '2026-05-22', is a '动态' (Update) announcement titled 'Aether 科研通用智能体亮相华为鲲鹏昇腾开发者大会' (Aether research general-purpose intelligent body debuts at Huawei Kunpeng Ascend Developer Conference), with a link to '阅读全文' (Read full text). At the bottom of the sidebar, there is a '展开更多动态' (Expand more news) button.

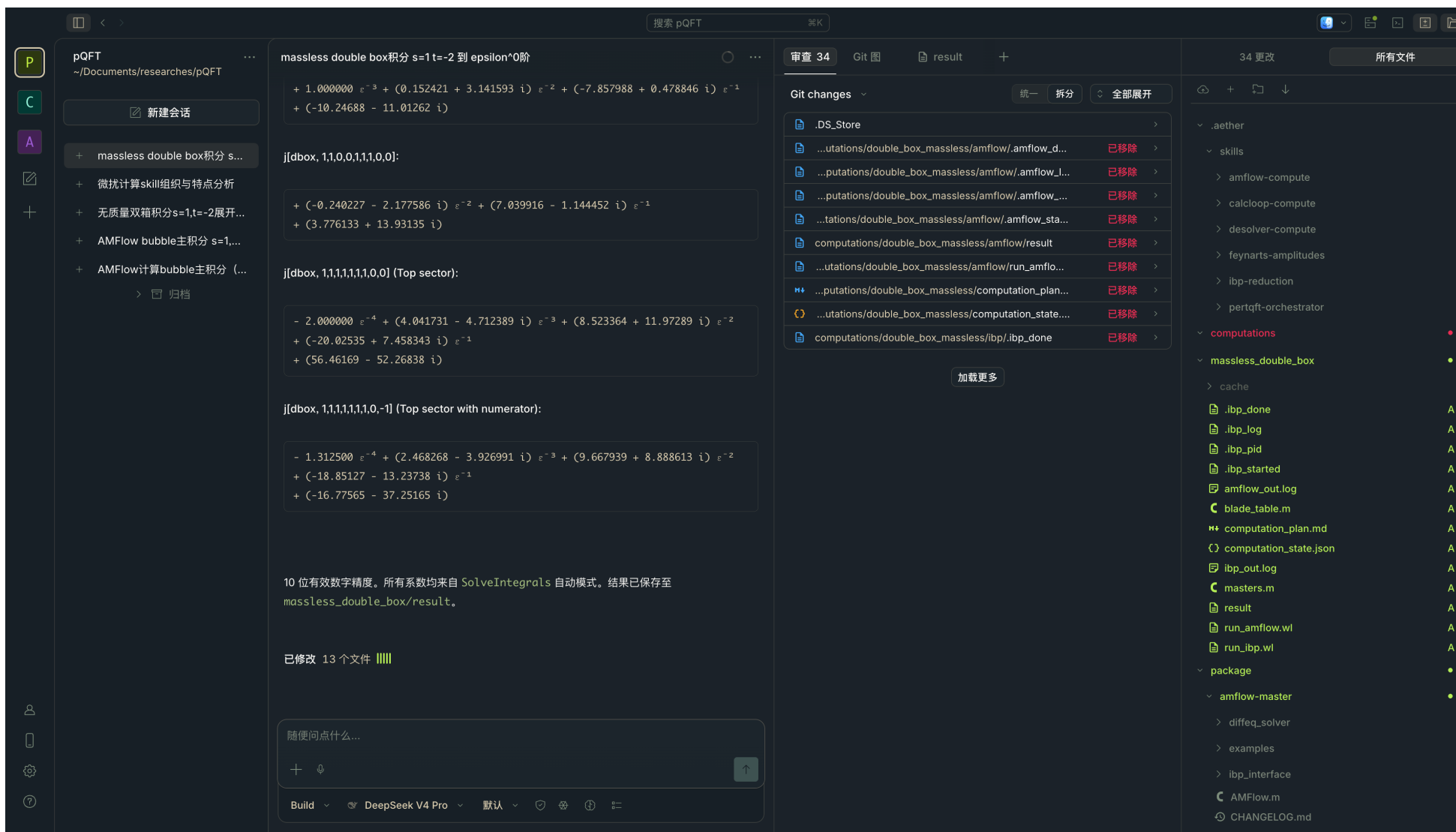
Highlights: inherited from OpenCode

- Open source
- Fully utilize the foundational capabilities of vibe coding



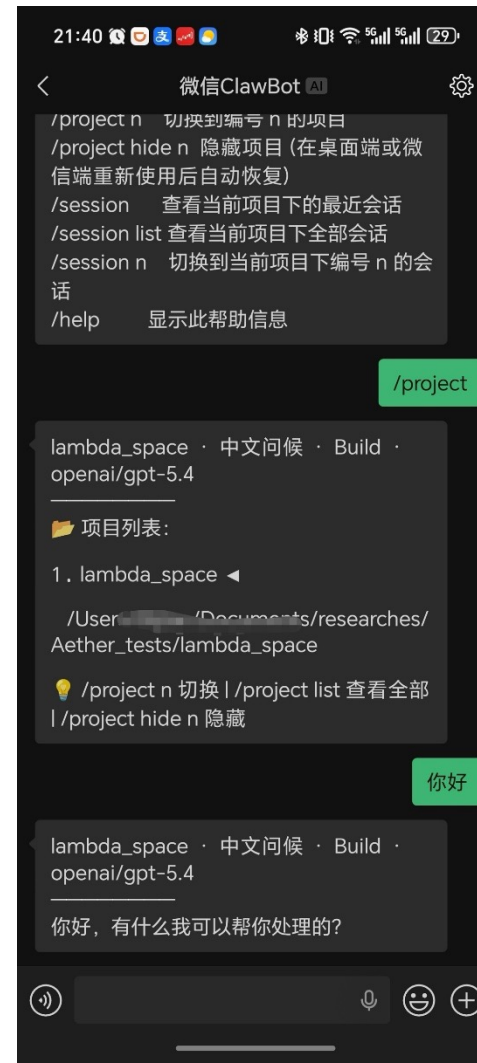
Highlights: highly enhanced UI

➤ User-friendly, ready to use out of the box



Highlights: mobile devices integration

- WeChat, QQ and FeiShu integrated, supporting remote work and voice input



Highlights: research adaptation

➤ Integrated professional skills/tools for scientific discovery (general & domain-specific)

The screenshot displays the 'Skill Market' website interface. At the top, there's a navigation bar with 'Skill Market' logo, '探索' (Explore), '我的 Skill' (My Skills), and '文档' (Docs). User options include '日间' (Day), '登录' (Login), '注册' (Register), '登出' (Logout), and '生态: Aether'. Below the navigation bar, the page title is 'SKILL MARKET / EXPLORE' and '发现可下载、可集成的 Agent Skill' (Discover downloadable, integrable Agent Skills). A subtitle reads '面向开发者的 Skill 目录。搜索、筛选并检查包内容后再下载集成。' (Skill directory for developers. Search, filter, and check package content before downloading and integrating). A search bar contains '搜索 Skill...' and filters show '全部分类' (All Categories), '全部' (All), '仅精选' (Only Curated), and '最新更新' (Latest Updates). The main content area features a grid of skill cards, each with a title, ID, description, author, and a '查看详情' (View Details) link. The skills listed are:

- scientific-brainstorming** (ID: scientific-brainstorming): Research ideation partner. Generate hypotheses, explore interdisciplinary connections, challenge... (Author: san zhang, 7 days ago)
- academic-plotting** (ID: academic-plotting): Generates publication-quality figures for ML papers from research context. Given a paper section or... (Author: san zhang, 8 days ago)
- academic-verify** (ID: academic-verify): Verify a research claim or academic citation by tracing it through publication → methodology → raw data →... (Author: san zhang, 4 days ago)
- scientific-critical-thinking** (ID: scientific-critical-thinking): Evaluate scientific claims and evidence quality. Use for assessing experimental design validity, identifying... (Author: san zhang, 3 days ago)
- citation-management** (ID: citation-management): Comprehensive citation management for academic research. Search Google Scholar and PubMed for... (Author: san zhang, 2 days ago)
- skill-creator** (ID: skill-creator): Create, edit, improve, or audit AgentSkills. Use when creating a new skill from scratch or when asked to... (Author: san zhang, 14 days ago)
- scientific-slides** (ID: scientific-slides): Build slide decks and presentations for research talks. Use this for making PowerPoint slides, conference... (Author: san zhang, 15 days ago)
- web-research** (ID: web-research): Searches multiple web sources, synthesizes findings, and produces cited research reports using delegated... (Author: san zhang, 15 days ago)
- read-arxiv-paper** (ID: read-arxiv-paper): Use this skill when when asked to read an arxiv paper given an arxiv URL (Author: san zhang, 15 days ago)

Highlights: research adaptation

➤ User-level memory mechanism; the Cron scheduler

Cron

查看定时任务、运行状态和最近运行历史。也可以用简短指令创建或修改定时任务。

全局 Cron 执行

启用或禁用自动执行。调度器会一直运行，以便任务定义和状态保持同步。

☐

调度器
运行中

执行状态
已启用

活跃任务
0/2

运行中
0

创建新的定时任务

生成

任务

共 2 个，0 个已过期

刷新

Daily memory reflection

direct / cron

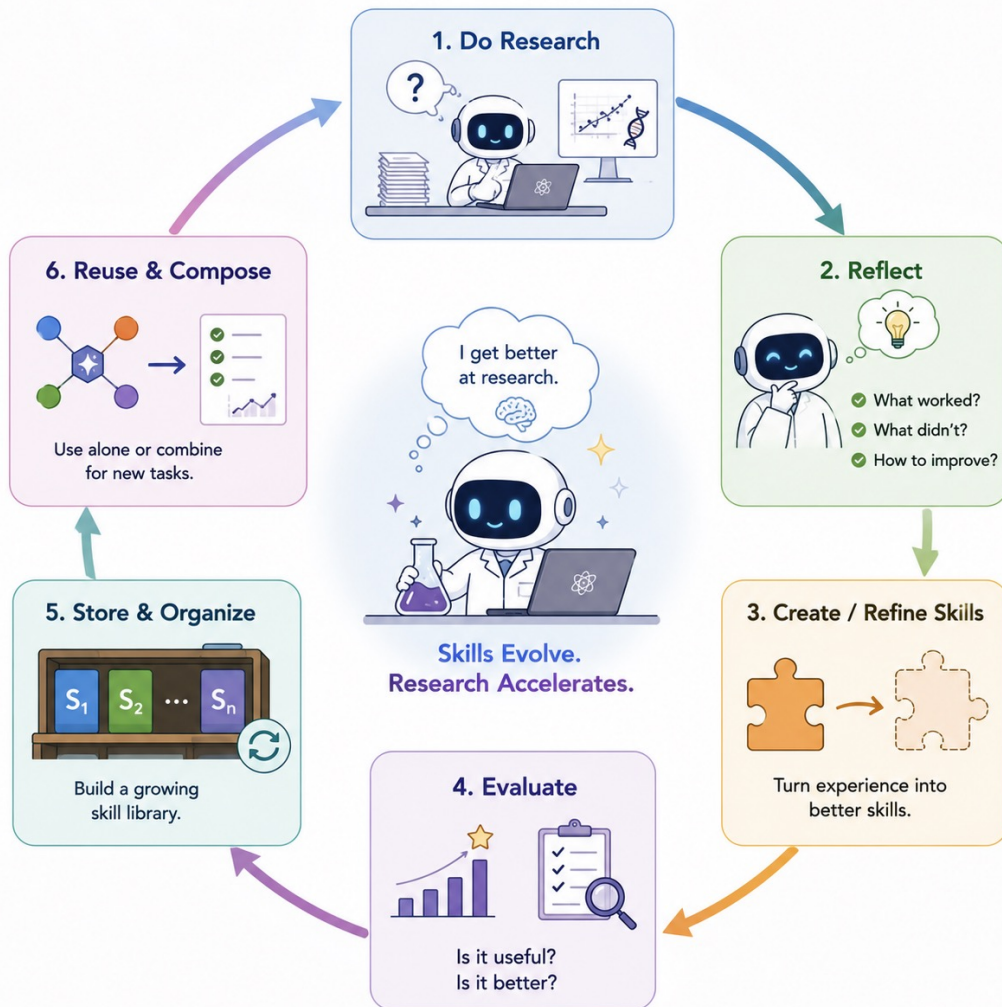
Paused

下次运行: 2026/6/1 03:00:00

上次运行: 2026/5/31 03:00:40

Highlights: research adaptation

➤ Self-evolving scientific research skills



Skill自进化

配置 AI 自动捕捉和优化可复用 Skills 的方式。

Skill 自进化

总开关。关闭后所有项目都不再触发后台 Skill 评审——下面的设置在重新打开前都不生效。

审查间隔

AI 每隔 N 次工具调用会自动回顾对话，将有价值的操作模式提炼为可复用 Skill。数值越小，提炼越频繁。

80

最大版本数

每个 Skill 保留的最大版本快照数，超出后自动清理旧版本。

100

单步字数上限

后台评审单步最多输出多少字符（单位 k，即千），超过就掐断这一步——防止模型在一步里停不下来狂吐。默认 300k（30 万）。

300

k

整场字数上限

一次后台评审整场最多输出多少字符（单位 k，即千），超过就停掉整场——防止评审一步步磨蹭停不下来。默认 1000k（100 万）。

1000

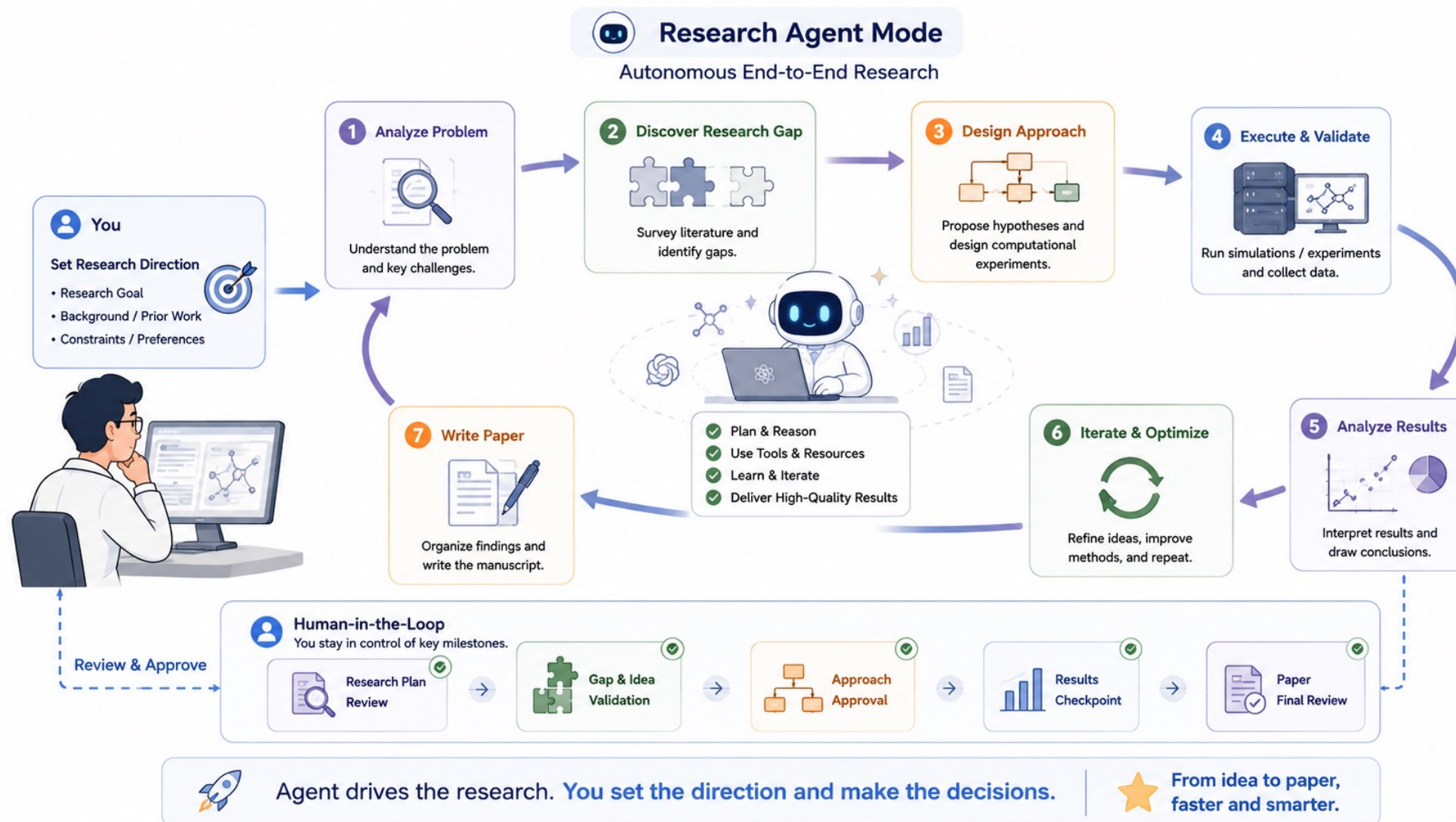
k

Skill 管家

定期统计 Skill 使用次数、把长期没用的标记为陈旧，并把 90 天没动过的 Skill 归档（归档副本可恢复）。关闭后停止所有自动维护。

Highlights: research adaptation

➤ Research agent mode (upcoming public release)



E.g.1: pQFT skills for high-energy physics



E.g.1: pQFT skills for high-energy physics

massless double box积分 $s=1$ $t=-2$ 到 ϵ^0 阶

我要算一个massless double box积分。动力学变量 $s = 1$, $t = -2$ 。展开到 ϵ^0 阶。

Pertqft-Orchestrator

Shell Check for existing computation state

```
$ find /Users/dsjian/Documents/researches/pQFT -name "computation_state.json" -maxdepth 3 2>/dev/null
```

问题 1 已回答

我将把此任务组织为本地计算 workflow：明确设置、构建计算计划、运行下游工具/skills，并组装最终结果。这是你期望的工作方式吗？

Yes, use the computation workflow (Recommended)

已探索 1 次读取

Now Phase 1: I need to clarify a few essential points before building the computation plan.

问题 3 已回答

你要计算的是哪种 double box 拓扑？

planar。该family所有的主积分

你想要哪个具体的积分？标量积分（所有propagator幂次=1）还是特定的master integral？

Full family master integrals

外部动量标记方式？

随便问点什么...

Build DeepSeek V4 Pro 默认

massless double box积分 $s=1$ $t=-2$ 到 ϵ^0 阶

$$+ (-0.240227 - 2.177586 i) \epsilon^{-2} + (7.039916 - 1.144452 i) \epsilon^{-1} + (3.776133 + 13.93135 i)$$

$j[\text{dbox}, 1,1,1,1,1,1,0,0]$ (Top sector):

$$- 2.000000 \epsilon^{-4} + (4.041731 - 4.712389 i) \epsilon^{-3} + (8.523364 + 11.97289 i) \epsilon^{-2} + (-20.02535 + 7.458343 i) \epsilon^{-1} + (56.46169 - 52.26838 i)$$

$j[\text{dbox}, 1,1,1,1,1,1,0,-1]$ (Top sector with numerator):

$$- 1.312500 \epsilon^{-4} + (2.468268 - 3.926991 i) \epsilon^{-3} + (9.667939 + 8.888613 i) \epsilon^{-2} + (-18.85127 - 13.23738 i) \epsilon^{-1} + (-16.77565 - 37.25165 i)$$

10 位有效数字精度。所有系数均来自 `SolveIntegrals` 自动模式。结果已保存至 `massless_double_box/result`。

Build · DeepSeek V4 Pro · 20分 56秒

已修改 13 个文件

随便问点什么...

Build DeepSeek V4 Pro 默认

E.g.2: research agent mode

费曼积分的FBI-DCT-NIS数值方法

研究计划确认 — FBI-DCT-NIS 联合费曼积分数值计算方法

前三个阶段已完成。以下是研究计划的摘要，请确认后再进入执行阶段：

研究问题 (4个可证伪问题)

ID	研究问题	可证伪标准
RQ1	FBI-DCT的 ϵ 展开流水线能否与NIS自适应重要性采样在统一数值框架中有效结合？	2个2圈基准积分上，30k次DCT评估内 $ \Delta /AMFlow < 1e-4$ ，方差缩减因子 $> 2\times$
RQ2	dct_nis_python代码库存在哪些架构/算法/鲁棒性不足？需要哪些改进？	识别 ≥ 3 个实质性缺陷，每个有量化预期收益

RQ3 FBI-DCT-NIS vs AMFlow ϵ^0 阶 $|\Delta|/AMFlow < 1e-3$ ， ϵ^1 阶 $< 1e-2$ 的逐 ϵ 阶精度/效率量化关系

RQ4 NIS精度天花板？RQSpline 30轮训练后 $\sigma/||\cdot|| < 1e-2$ 且 RQSpline改进 $< 2\times$ 则已达上限

可验证的Claims

- 可行性与精度: FBI-DCT-NIS在2圈积分上可行，方差缩减 $> 2\times$ ， ϵ^0 阶精度 $< 1e-4$
- 代码审查: dct_nis_python存在 ≥ 3 个实质性缺陷，改进后可支撑实用化
- 精度对比: ϵ^0 阶 $|\Delta|/AMFlow < 1e-3$ ，与成熟工具可比
- 误差量化: NIS统计误差与DCT截断误差可通过误差传播公式分离

Outline

I. AI for established fields: current status

II. AI for frontier fields: challenges

- a. Complexity of frontier research tasks
- b. The fundamental research–coding gap
- c. Accessibility for the broader research community

III. Aether: an integrated vibe research platform

IV. Summary and outlook

Summary and outlook

- AI for scientific discovery is evolving from isolated algorithmic tools toward integrated research platforms
- The next generation of AI research agent should augment the entire scientific discovery process, rather than individual research tasks
- Because scientific research fundamentally differs from coding, dedicated research agents—not generic coding agents—are essential for effective AI-assisted scientific discovery
- General and domain-specific research skills will continue to be developed by experts, ultimately becoming valuable training resources for future foundation models

Thank you!