SciencePedia

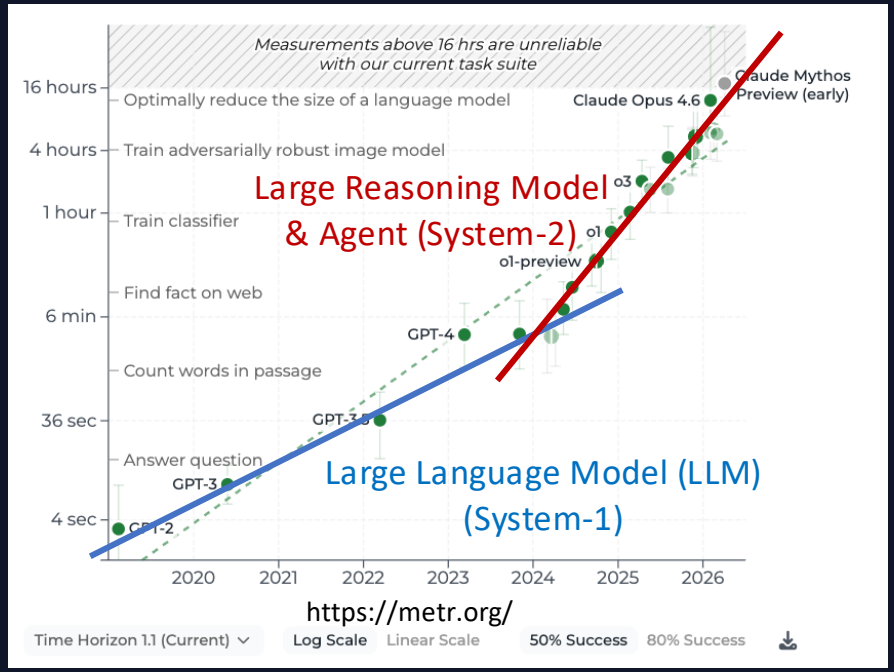
从大语言模型到大知识模型

Agent赋能前沿科学发现的形式化路径

陈 锟　中国科学院理论物理研究所

AI智能体赋能核物理研究研讨会 · 2026 年 7 月 4 日

Agentic Science 现状：从工具到研究主体



范式转变：从 AI as Tool（计算工具）到 AI as Reasoning Subject（推理主体）

AGI 让世界跑快 1000 倍 —— 然后？

每一次大加速都伴随基础设施重构 —— bottleneck migration

速度革命	解决旧瓶颈	出现新瓶颈	应运而生的基础设施
工业革命 *18-19 c.*	**体力**劳动	资本跨地流转	**股份公司** · 银行 · 专利法
计算革命 *1940s-*	算术	**数据跨机器流**转	**文件系**统 · **数据**库 · TCP/IP
互联网革命 *1990s-*	信息传播	**注意力分配**	**搜索引擎** · Wikipedia · GitHub
AGI 革命 *now-*	**个体推理**	**???**	**???**

下一页 AGI 时代的「新瓶颈 + **新基**础设施」该长什么样？我们用一页给出答案 —— **标准化的「科学交流形式」是什么？**

编程已被工业化，下一个被工业化的将是科学研究本身

迭代节奏正在从“年”压缩到“天”

维度	手工作坊时代	科研工业化时代
生产模式	个人 / 小团队手工	标准化流水线 + 自动化
协作节奏	年度学会 + 书信	实时协作 + 持续集成
知识量级	数百篇 / 年	数百万篇 / 年
进步驱动力	个体天赋	基础设施 + 系统工程

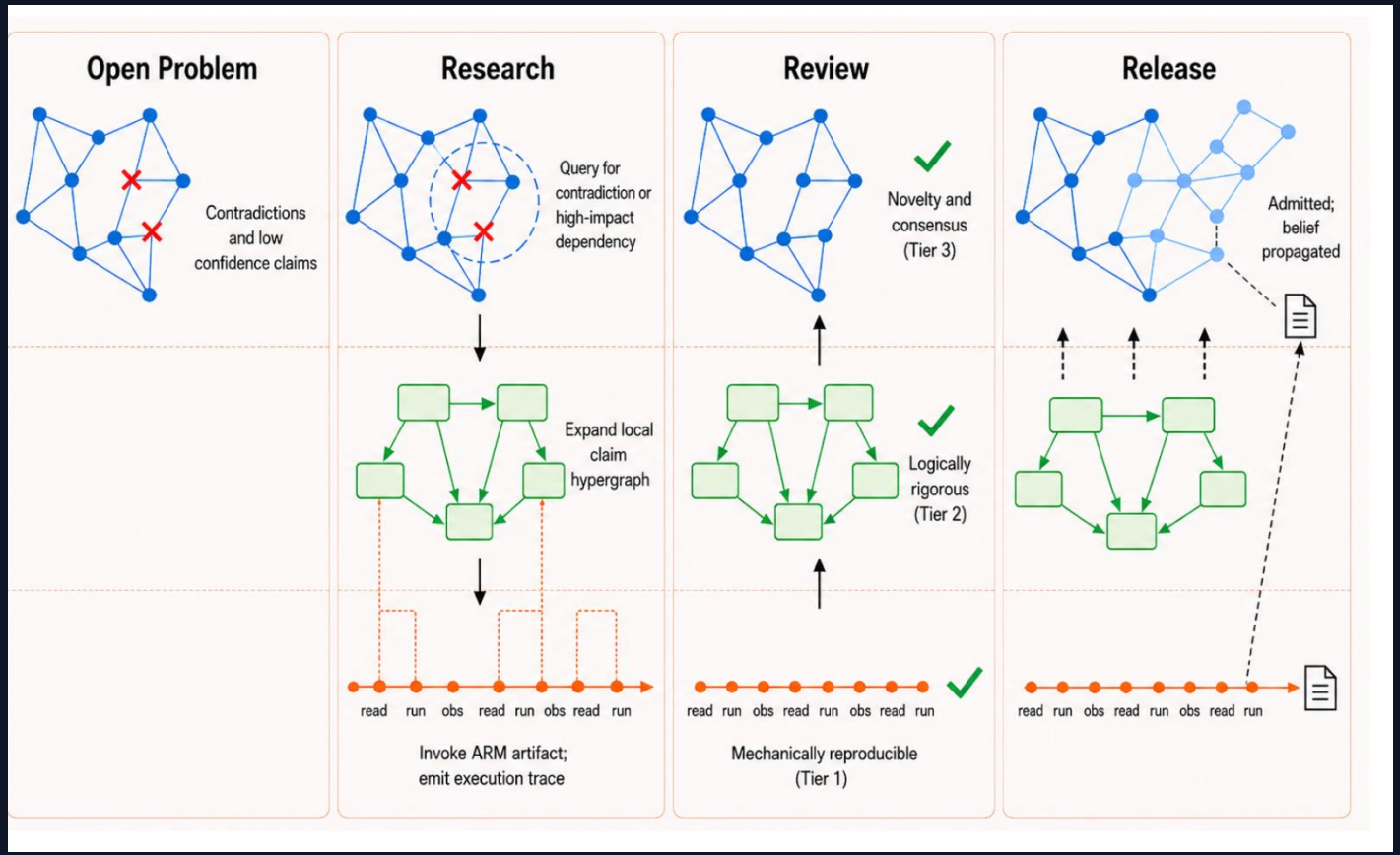
> *1660 年皇家学会确立的协议——同行评议、论文格式——300 余年未曾根本更新。*
> *科研工业化的基础设施，现在才刚开始被建造。*

我们的探索：Scientific World Model

大知识模型
（LKM）

似然推理
形式化
（Gaia）

人类/AI
思维链/论文
Agentic-Ready Manuscript
(ARM)



LKM：机器可读的科学知识共同体

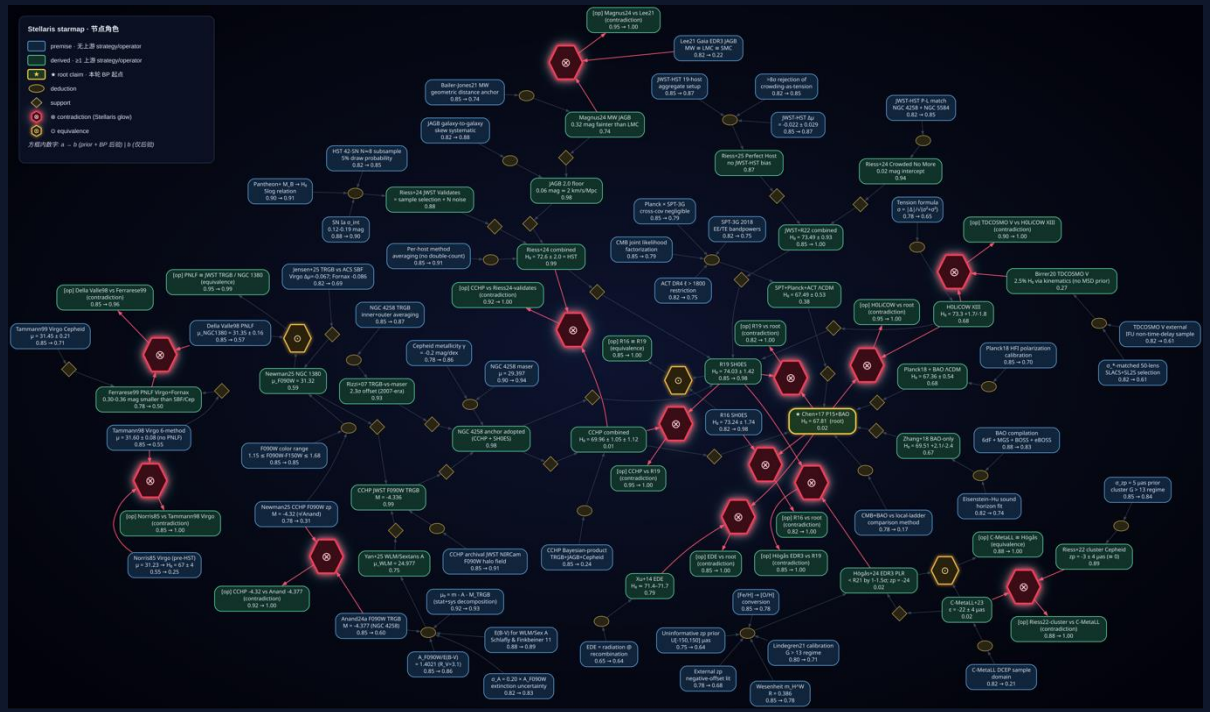
人类科学家 → arXiv / PubMed = AI Agent → LKM

	传统知识图谱	LKM
基本单元	实体-关系三元组	带来源、置信度、推理链的科学断言
来源	手工标注 / 规则抽取	从论文语义自动提取
推理	路径查找	推理链
矛盾处理	通常没有	显式标记并追踪
知识演化	静态或半静态	随新论文实时更新

~10 亿科学命题 · 全部主要学科 · 通用命题级语义搜索引擎

大知识模型 = 科学命题为节点的推理网络 (前置命题 → 结论)

LKM 让矛盾变得可见、可追踪、可量化



哈勃常数危机

本地 vs CMB-BAO H_0

SH0ES ~73 vs Planck ~67.8，>4σ

TRGB vs Cepheid 校准

TRGB 系统低于 Cepheid，零点分歧

JAGB MW vs LMC 通用性

跨星系迁移性存疑

PNLF 系统差

Virgo vs Fornax 距离偏差

Virgo 距离尺度修订

pre-HST → Tammann 反复修正

大知识模型 =科学命题为节点的推理网络 (前置命题 → 结论)

ARPES vs DFT/多体重整化

实验 vs 理论有效质量分歧

Fröhlich vs acoustic/defect

迁移率瓶颈归因争议

phonon/Urbach/linewidth 能标

各组特征能量互相矛盾

Single vs multi-channel PL

温度淬灭模型选择分歧

纳米晶 vs 体相能否互通

量子点规律难外推到 bulk

材料第一性原理方法与实验的系统偏差

矛盾是科学进步的前沿

LKM 让矛盾变得可见、可追踪、可量化

案例一：哈勃常数危机

67.4 km/s/Mpc
CMB 测量

VS

73.0 km/s/Mpc
造父变星

4–5σ 显著性矛盾，上千篇论文争论

LKM：实时追踪完整证据链，给每个测量方法标注置信度

案例二：DFT 系统偏差

- DFT 对带隙的预测存在系统性低估
- LDA / GGA / HSE 泛函给出相互矛盾数值
- LKM 给出跨方法的置信度分布

矛盾 = 科学的前沿 —— 传统做法：人工梳理 → LKM：矛盾被结构化标注，可直接查询

应用举例：循证分析

结论是否有证据支持，定位争议点和潜在矛盾，下一步研究方向

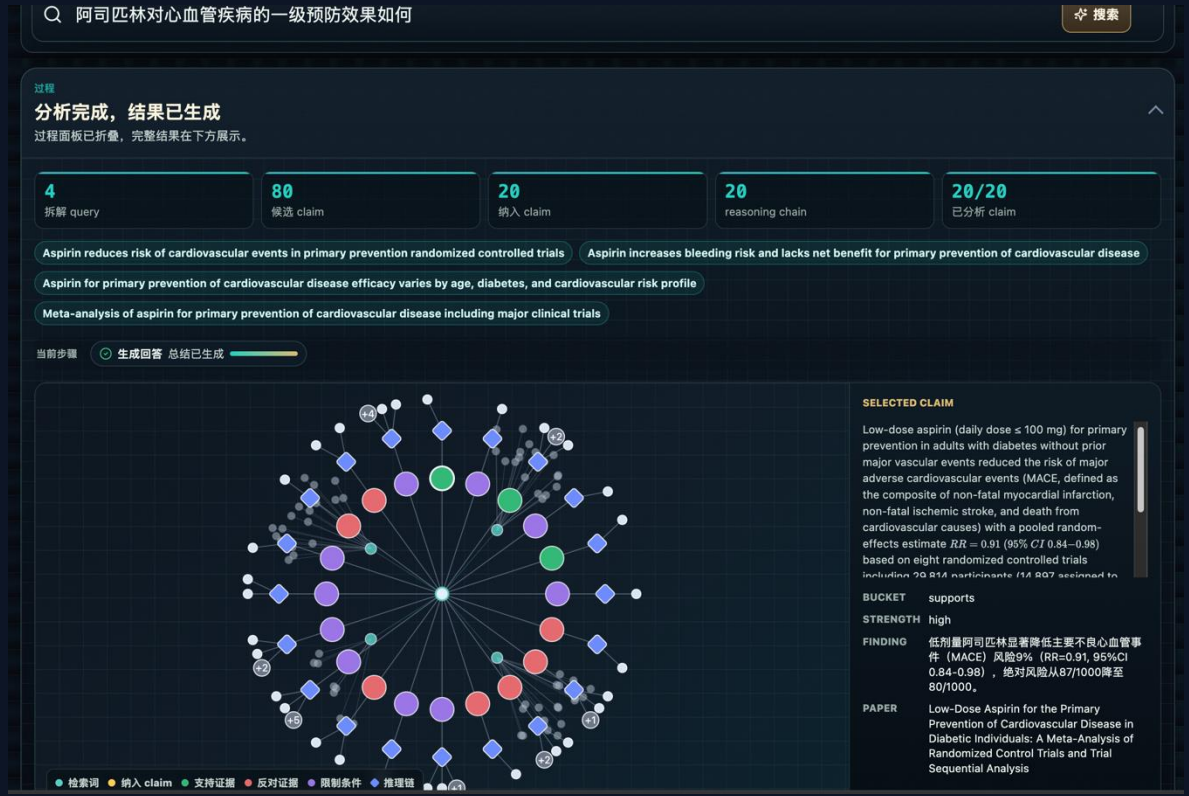
循证分析工作流

1. 从一个结论出发：这个 claim 是否可靠？

2. 向上追溯：它依赖哪些观测、假设、校准和模型？

3. 横向比较：是否存在独立证据、等价证据或共享数据源？

最终产物不是摘要，而是一份可复查的 evidence dossier。



循证分析的价值：把专家脑中的文献判断，变成 agent 可以追踪、比较、复用的结构

应用举例：循证分析

结论是否有证据支持，定位争议点和潜在矛盾，下一步研究方向

循证分析工作流

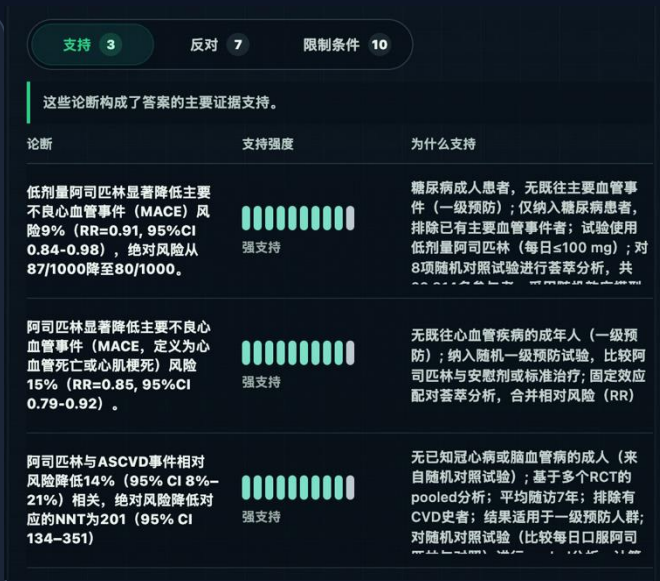
1. 从一个结论出发：这个 claim 是否可靠？

2. 向上追溯：它依赖哪些观测、假设、校准和模型？

3. 横向比较：是否存在独立证据、等价证据或共享数据源？

4. 向下展开：如果它错了，哪些下游结论会被影响？

最终产物不是摘要，而是一份可复查的 evidence dossier。



循证分析的价值：把专家脑中的文献判断，变成 agent 可以追踪、比较、复用的结构

应用举例：循证分析

结论是否有证据支持，定位争议点和潜在矛盾，下一步研究方向

矛盾诊断

定位证据方向不一致的原因。

高	糖尿病亚组中MACE获益结论不一致	支持证据(claim 2)显示低剂量阿司匹林显著降低糖尿病患者MACE风险9%（RR=0.91, 95%CI 0.84-0.98），而反对证据(claim 15)显示无显著降低（RR=0.90, 95%CI 0.78-1.05, p=0.17）。两个meta分析纳入的试验集合不同，导致结论分歧，可能源于纳入试验的基线风险或随访时长差异。
高	非致死性心梗与死亡率结局不一致	支持证据(claim 4)显示阿司匹林降低MACE（心血管死亡或心梗）风险15%（RR=0.85, 95%CI 0.79-0.92），而反对证据(claim 16)显示心血管死亡率无显著降低（RR=0.99, 95%CI 0.90-1.08）。这种差异提示阿司匹林主要减少非致死性心梗事件，但对心血管死亡无影响，导致总体净获益不确定。
中	女性亚组与混合人群效果差异	支持证据(claim 10)在混合人群中显示ASCVD事件降低14%（NNT=201），而反对证据(claim 12)在女性健康研究（隔日100mg）中未达到统计学显著（RR=0.91, 95%CI 0.80-1.03, p=0.13）。这一矛盾可能源于性别差异或用药频率（隔日 vs 每日）的影响。

空白与下一步

把限制条件转化为后续验证工作。

验证 近期

糖尿病一级预防中阿司匹林净获益是否可靠？不同meta分析结果为何冲突？

→ 开展统一纳入标准的个体参与者数据荟萃分析，按糖尿病病程和基线风险分层

推荐追问

阿司匹林糖尿病一级预防效果是否存在净获益？

追问这个缺口

糖尿病亚组中MACE获益结论不一致

方法 近期

阿司匹林对心血管死亡无显著影响的结论是否受剂量、随访时间或终点定义影响？

→ 评估不同试验中心血管死亡定义的一致性，并分析剂量与随访时长的交互作用

推荐追问

低剂量阿司匹林能否降低一级预防人群的心血管死亡率？

追问这个缺口

非致死性心梗与死亡率结局不一致

循证分析的价值：把专家脑中的文献判断，变成 agent 可以追踪、比较、复用的结构

理论基础： Cox 定理 - 似然逻辑必然导向概率

在此基础上发展出Gaia科学形式化语言

科学命题 → 逻辑关系 → 一致性公理 → Cox 定理 → **概率论**

Cox 定理（1946）三条公理

① **可比性**: 任意两命题的可信度可比较排序
② 单调性: $A \supset B \Longrightarrow P(A) \leq P(B)$
③ 复合一致性: $P(A \wedge B)$ 由 $P(A)$ 与 $P(B|A)$ 唯一确定

▶ **推论: 形式化语义正确 → 可信度度量客观唯一**

谓词逻辑关系 → 传统知识图谱
三元组

命题逻辑关系 → 似然推理图
蕴含 / 合取 / 析取 / 矛盾 / 等价

关联关系 → 贝叶斯网络
误差模型 → 实验观测

因果关系 → 因果推断
Pearl 三阶梯: 关联 → 干预 → 反事实

把论文里的命题抽出来、理清它们的逻辑关系——
*这件看似很"**工程**"**的事，在数学上等价于**对整个科学知识体系做客观贝叶斯重构。*

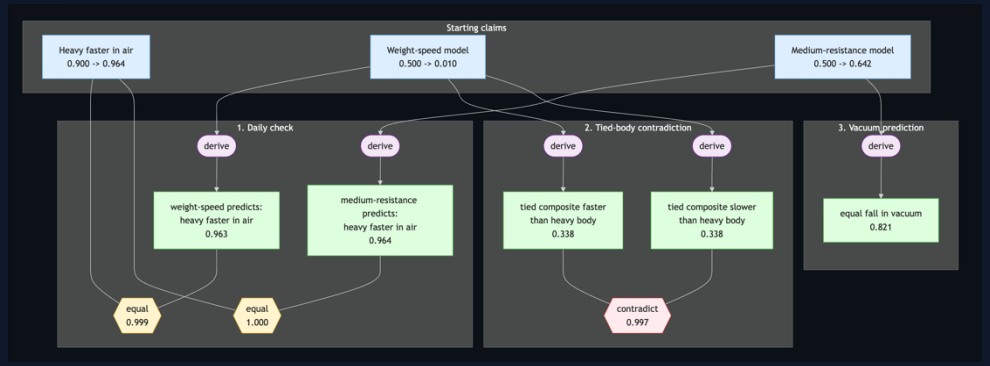
Gaia 如何把矛盾变成可计算的命题网络

从「有矛盾」到「机器如何量化矛盾」

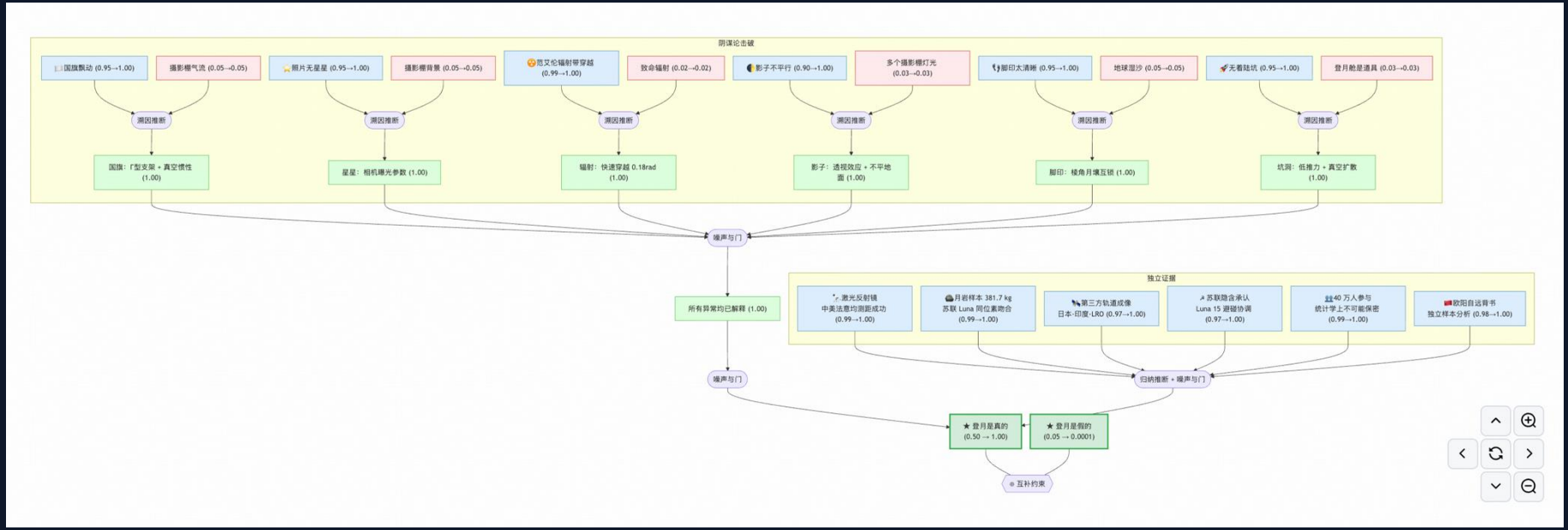
伽利略比萨斜塔思想实验的 Gaia 形式化

```python
from gaia.lang import (
    claim, deduction, abduction, contradiction,
)

# 观测与两种竞争解释
obs       = claim("重物在空气中比轻物下落更快。")
aristotle = claim("速度正比于重量—重者更快。")
air_drag  = claim("速度差源于空气阻力，不是重量。")

abduction(observation=obs,
          hypothesis=air_drag,
          alternative=aristotle,
          reason="两者都能解释日常现象")

# 亚里士多德定律的自相矛盾
slower = claim("复合体比重球慢")
faster = claim("复合体比重球快")
deduction([aristotle], slower,
          reason="轻球拖拽重球")
deduction([aristotle], faster,
          reason="总重量更大")

contradiction(slower, faster,
    reason="同一复合体不能同时更快又更慢")
```

1. **搜集**证据链
2. **自**动标记：两条链是否有**共享前提（是否真的独立）**
3. 计算矛盾的贝叶斯证据比
4. 输出：这是一个真实的、**高置信度的科学矛盾**



登月是不是假的？--- Gaia形式化后的推理图



结论

12 条证据线 → 阴谋解释需同时维持 6 个独立造假 + 多国航天机构配合 + 40 万人集体沉默。交集概率趋近于零。登月的真实性不是一个观点 — 它是数学必然。

登月是不是假的？

Gaia 形式化推理 · 78 节点 · 26 策略 · Junction Tree 精确推断

✅ 登月是真的	❌ 登月是假的
先验 50% → 后验 **99.9999%**	先验 5% → 后验 **0.01%**

第一条线 ▸ 六大「异常」的科学解释

国旗飘动	「型支架 + 真空惯性振荡	1.00 / 0.05
照片无星星	日照面曝光参数（~1/250s f/5.6）	1.00 / 0.05
范艾伦辐射带	快速穿越 · 总剂量仅 0.18 rad	1.00 / 0.02
影子不平行	广角透视 + 不平月面	1.00 / 0.03
脚印太清晰	棱角月壤颗粒互锁	1.00 / 0.05
无着陆坑	低推力（~1400 lbf）+ 真空扩散	1.00 / 0.03

第二条线 ▸ 六条独立证据

激光反射镜	阿波罗部署 · 中美法意均测距成功	1.00
月岩样本	381.7 kg · 苏联 Luna 同位素吻合	1.00
第三方成像	日本 SELENE · 印度 Chandrayaan · LRO	1.00
苏联承认	冷战对手独立追踪 · 从未否认	1.00
规模论证	40 万人 · Grimes 模型 < 4 年暴露	1.00
中国验证	欧阳自远背书 · 独立样本+测距	1.00

github.com/huangy22/moon-landing-hoax-gaia

用 Gaia 形式化真实论文 · 自动产出推理图 + 计算信息增益

论文 → Gaia 知识包 → 推理结构机械校验 → 作为子图注入 LKM

案例
arXiv:2512.19382

1 论文
arXiv:2512.19382
Ï Þ Ó Ñ Ø Ñ Ò Ô Ñ Þ Ñ Þ Ò Ø Ò Ø Ř Ò Ð
Ĝ Ø Ñ Ñ Þ Ø Ð Ô Ï Ø Þ Ô Ñ Œ

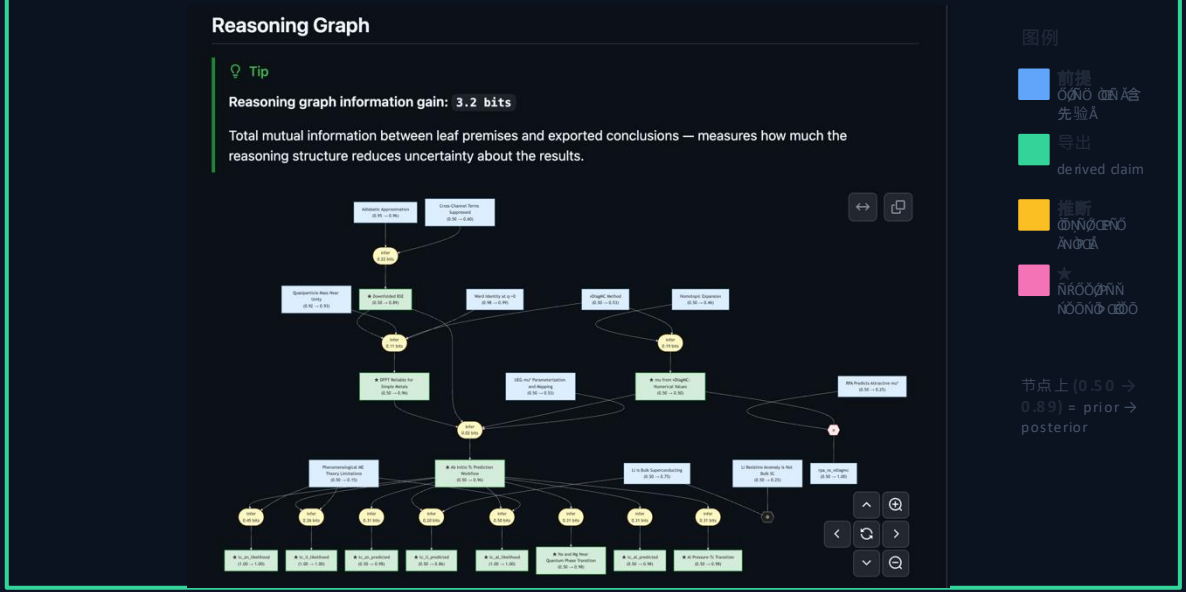
2 ĠMÒM形式化
SuperconductivityElectronLiquids.gaia
claim + derive + 形式化
66 commits · 9 branches

3 推理图 + 信息增益
前提 ⊓ 推断 ⊓ 结论
信息增益 = 3.2 bits ★

4 注入 LKM 子图
作为可寻址子图加入
LKM 全局命题网
供其它 agent 复用

Ī Ñ M Ð Ð Ò Ò Ð Ñ Ġ Ø M Ő O ï 自动生成自
Ï Þ Ő Ñ Ø Ñ Ò Ô Ñ Þ Ñ Þ Ò Ø Ò Ø Ř Ĝ Ô Ñ Ñ Þ Ø Ð Ô Ï Ø Þ Ô Ñ Œ Ñ M Ò M

信息增益 **3.2 bits**



信息增益 Ċ ĐĚ Ñ Ò Œ Ě 叶节点前提与导出结论之间的互信息 àà 量化「这篇论文的推理结构本身给了 ı Ħ ì 多少不确定性降低」。
每个 Ð Ñ M Ò M 包都自带这个数 ⊓ **ı Ħ ì 知道哪些论文真的「增加了知识」**，哪些只是重述既有内容。

梳理人类知识：从论文 → 标准化科研 workflow

LKM 不止存结论，更把跨论文、可复现的「方法配方」抽成标准 workflow（输入 → 方法 → 验证 → 输出量），从而可比较、可基准化

机器学习势函数	å t ǐ i Ń	能带 · 带隙		过渡态 · 能垒	ǐ
MLIP 训练 / 验证	DFPT	DFT · GW	CSP	NEB	BTE

…… *每类都有标准输入 / 方法 / 验证 / 输出量*

示例 **电声耦合 λ 基准（DFPT）· Harbor/ARM format**

标准 workflow：晶体结构 → DFT 自洽 → DFPT（声子 + 电声矩阵元）→ **α²F(ω)** → **λ, ω_log** → **Tc**（McMillan–Allen-Dynes）

两层任务：**L1** υ f ĥ Ğ Ă Å ' q Ĉ Ç r tf **L3** 由结构经 G Ğ ĩ i 算 '（Ĉ Ď Ç 即用 C Ĉ Ĉ Č 待构建）

规模与判分：276 参考点 · 123 篇论文 · 5 类材料；金标 = 论文报告的第一性原理 λ（容差 ≤15% ≤15%，与实验中位 ~9% 符合）

github.com/kunyuan/dfpt-lambda-benchmark

把一篇篇论文里的「怎么算」，沉淀成可复现、可基准、可比较的标准 R Ŏ Ø Ď Ņ Ď R C Ė Ī ì f f 这是 I Ħ ì 给科学知识做的结构化

First-Principle Agentic Science-材料科学中的一个Case Study

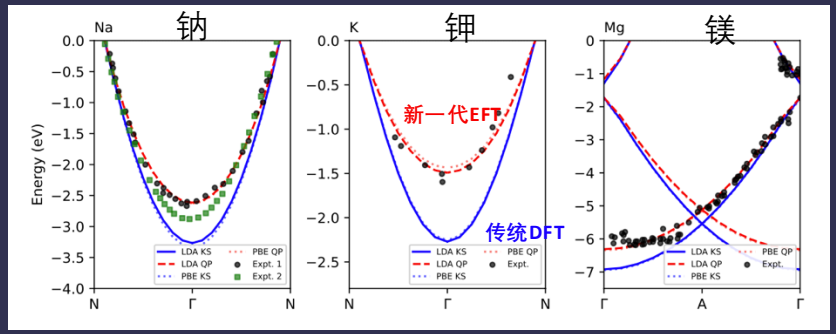
基于有效场论(EFT)重建Kohn-Sham电子结构理论的第一性原理基础

60年悬而未决的基础问题

Kohn-Sham(KS)理论是全球计算材料学的基石。

但KS特征值为何能代表准粒子能量，60年来没有第一性原理的回答。

碱金属带宽误差20-35%，DFT+DMFT的double counting问题，根源都在于此。



AI协作下的基础理论突破

从有效场论(EFT)出发，第一性原理推理：

KS Hamiltonian = 准粒子的leading-order运动方程

导出解析修正公式 | 7种元素验证 | 定量符合实验 | 计算<1秒

AI主导7种元素系统数值验证

元素	Li	Na	K	Ca	Mg	Al	Si
修正	31%	22%	36%	18%	9%	4%	2%
ΔE(Ha)	2.63	2.70	1.72	2.51	4.07	5.69	7.57

碱金属修正大(20-35%) → 解决ARPES问题 | Al/Si修正小(<5%) → 解释DFT为何已足够

1 精准材料筛选

修正公式嵌入高通量pipeline，
金属材料零成本提升精度20-35%

2 攻克强关联材料

d/f电子体系的统一理论通道，
系统消除double counting

3 下一代计算方法

为GW/DMFT提供有物理意义的起点，
替代60年经验拟合路线

4 超导材料发现

更精确电声耦合 → 更可靠Tc预测
高温超导体的机理与第一性原理计算

First-Principle Agentic Science: 一种突破前沿科学难题的新范式

从物理直觉到理论推导、数值验证、论文撰写：AI Agent参与全流程

全链条自动化研究闭环

LKM：通过搜索矛盾对寻找理论/实验 gap
提出EFT思路，识别KS理论的根本问题

AI Agent：批量尝试基础理论创新
推导公式、验证数学一致性、
尝试多种理论方案

AI Agent：大型计算软件定制化
编写7种元素的完整计算pipeline
(材料第一性原理计算软件, 原子求解器, QP修正)

AI Agent：迅速铺开高通量计算
跑DFT+ EFT修正, 与实验/eDMFT对比
系统验证across所有元素

批量close 理论/实验gap， 消除LKM内部矛盾
检验物理图像, 调整叙事, 提出新方向

效率革命

	传统模式	人+AI Agent
团队规模	3-5人课题组	1人 + Claude Code
研发周期	1-2年	数周
元素覆盖	1-2种	7元素分钟级验证
代码规模	团队分工开发	AI生成+人审核

Agentic Science At Scale

不是一个结果，是建立一个可持续产出突破的全自动闭环

可复制：同一范式可推广到量子化学、核物理、天体物理等理论密集领域

这不是 "AI猜出了一种新的材料构型"

而是 "AI帮助找到了新的理论框架、实现了新的代码、给出了新的实验预言"

测试用例：First-Principle Agentic Science

让 AI Agent 参与从理论推导到数值验证的完整第一性原理研究

Agent 闭环：**物理直觉** → **理论推导 · EFT** → **写代码** → **数值验证** → **对齐实验** ↻ **迭代**

示例 QP-EFT Agent Benchmark · 让 Agent 真做第一性原理

问题：裸 KS（DFTK）把碱金属 ARPES 带宽高估 **20–35%**；EFT 冻芯修正 **E_QP-E_F = z_core·(E_KS-E_F)** 对齐 ARPES ~0.1 eV

任务：写 **run_qp.py**ŗ GĞİ Ħ Π Ï Ĩ FǏ IJŗ ĦǏ

三层（信息递减）：**L1** 给公式+ 形状因子表 · **L2** 自己算形状因子 · **L3** 只给物理设定，自己从 EFT 推 z_core

评测：**公开** Í MŒ Õ**开**发，隐藏 ĦÒ Ŋ 评测；**最近**带 Ī Ì Ǐ Ĝ：ĦÊ ĆĒĈĐ CÌ ŊÊ ĆĒĈĈ ÑIJ（ŊŎŐŇ ĆĒĈĊEĆĆĒĈĐĐ，总体 ĆĒĈĎĊ ÑIJ）fí 含反作弊

github.com/kunyuan/qp-eft-agent-benchmark

「AI **推出新理**论框架、写出代码、给出可验证预言」—— **而且能被客**观考评

First-Principles 在自然科学的覆盖图

这套范式能管多宽 —— 共享「first-principles + 控制精度 + 跨系统泛化」三件套的开放问题

凝聚态

- ARPES bandwidth ★ PRL prototype
- 强关联 d/f 双重计数
- 非常规超导 pairing

Tools DFT · EFT · DiagMC · DMFT

核 / 强子物理

- 核力 from QCD
- 核素 binding energies
- 强子谱 hadron spectrum

Tools chiral EFT · GFMC · IM-SRG

天体物理

- 中子星 EoS（GW 约束）
- 致密天体内部结构

Tools 核 EFT + GR

材料 / 化学

- 反应速率常数
- 材料能隙 / 输运
- 催化机制

Tools CCSD(T) · DMRG · DFT+

精密 QED

- g-2 异常磁矩
- Casimir force precision
- 高阶辐射修正

Tools QED 微扰展开

Lattice QCD

- 强子谱
- 矩阵元 / form factors
- 有限温物质 / EoS

Tools lattice QCD ab initio

核心判断

1 Agent 已不再是工具——它是科研主体的雏形，增长看不到天花板

2 科研工业化缺一块：机器可读的科学知识共同体

3 I ĦÌ 存储、ĠMÒM推理、ĖĪÌ 形成Ė ŊÑŌP时代科学知识累积的飞轮

4 F ŎŔ 定理保证：科学命题一旦形式化，概率客观且唯一

5 First-principles Agentic Science: 一种Agent突破自然科学难题的新范式

工业革命时代，决定未来的不是谁先用上蒸汽机，而是谁先建起铁路网络。