上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

大模型的通用科学能力边界

Probing Scientific General Intelligence of LLMs with Scientist-Aligned Workflows

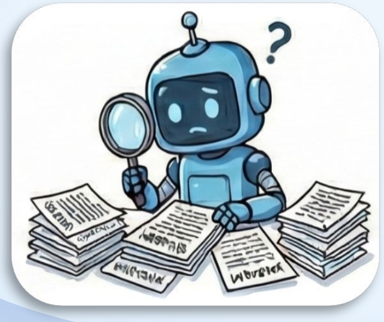
演讲人：徐望瀚

From AGI to SGI（通用科学智能）

AI能够胜任很多日常工作，但在科研中的环节中，**能力缺乏**。

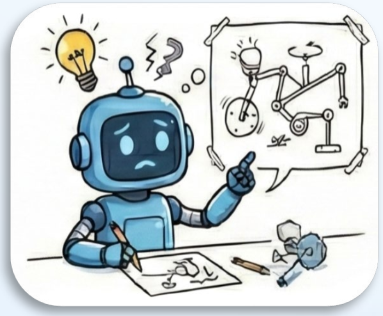
通用生活场景

知识问答，日常聊天…



文献调研

深度调研精度低



创意生成

创意生成可行性差

代码编写

干实验稳定性低

实验操作

湿实验动作错误

实验分析

数据分析推理能力弱

如何定义SGI（通用科学智能）？

Critical inquiry in a text-based environment: Computer conferencing in higher education

DR Garrison, T Anderson, W Archer

The internet and higher education, 1999 · Elsevier

The purpose of this study is to provide conceptual order and a tool for the use of computer-mediated communication (CMC) and computer conferencing in supporting an educational experience. Central to the study introduced here is a model of community inquiry that constitutes three elements essential to an educational transaction—cognitive presence, social presence, and teaching presence. Indicators (key words/phrases) for each of the three elements emerged from the analysis of computer-conferencing

SHOW MORE

1.Deliberation（思辨）

分析证据并进行批判推理

2.Conception（构思）

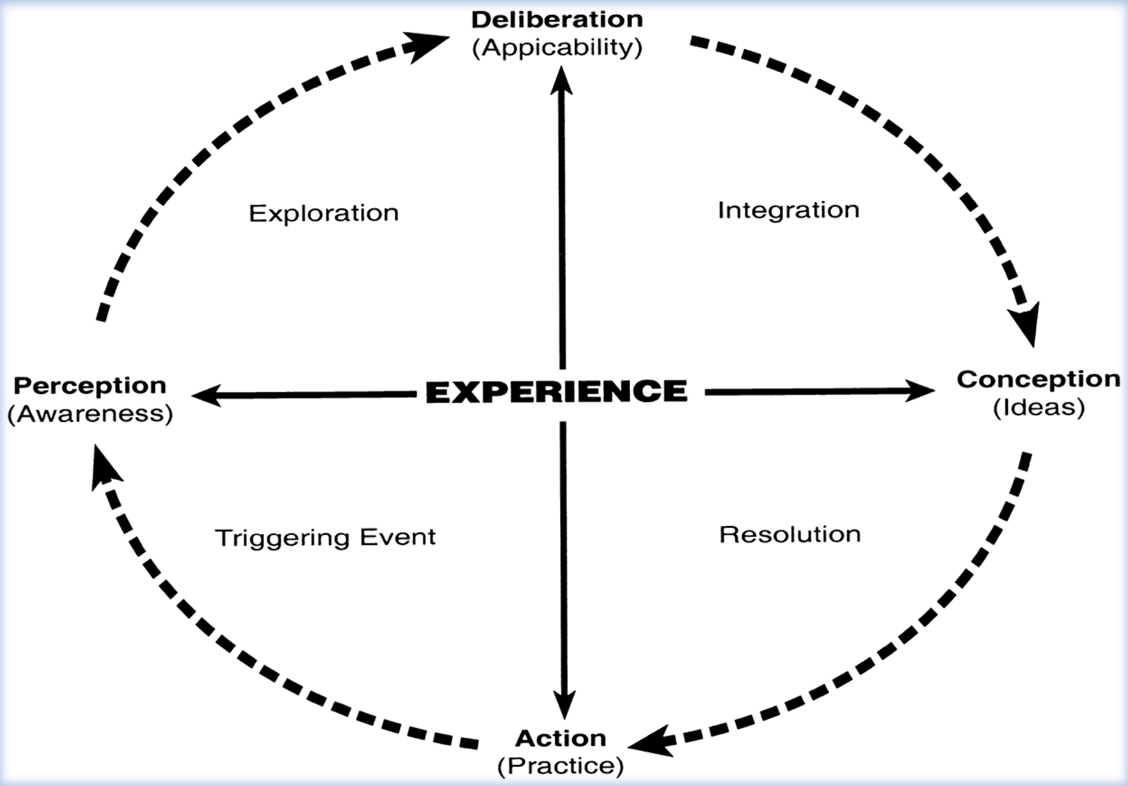
提出新思路和方法

3.Action（行动）

执行实验或操作验证假设

4.Perception（感知）

观察结果并更新认知

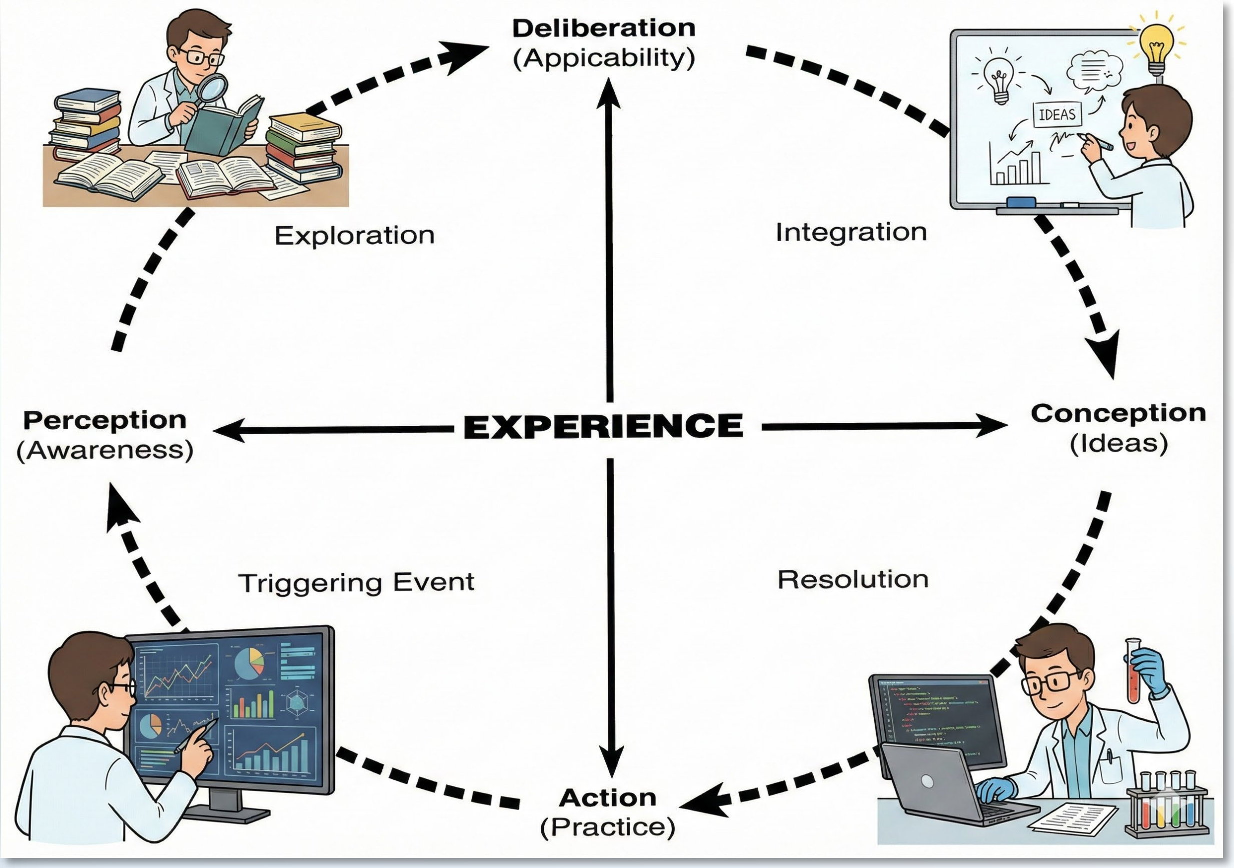


Practical Inquiry Model（PIM）
实践探索模型

与真实科研场景对齐

Deliberation（思辨）
文献，资料等调研分析

Perception（感知）
实验结果，数据分析



Conception（构思）
思考新的idea

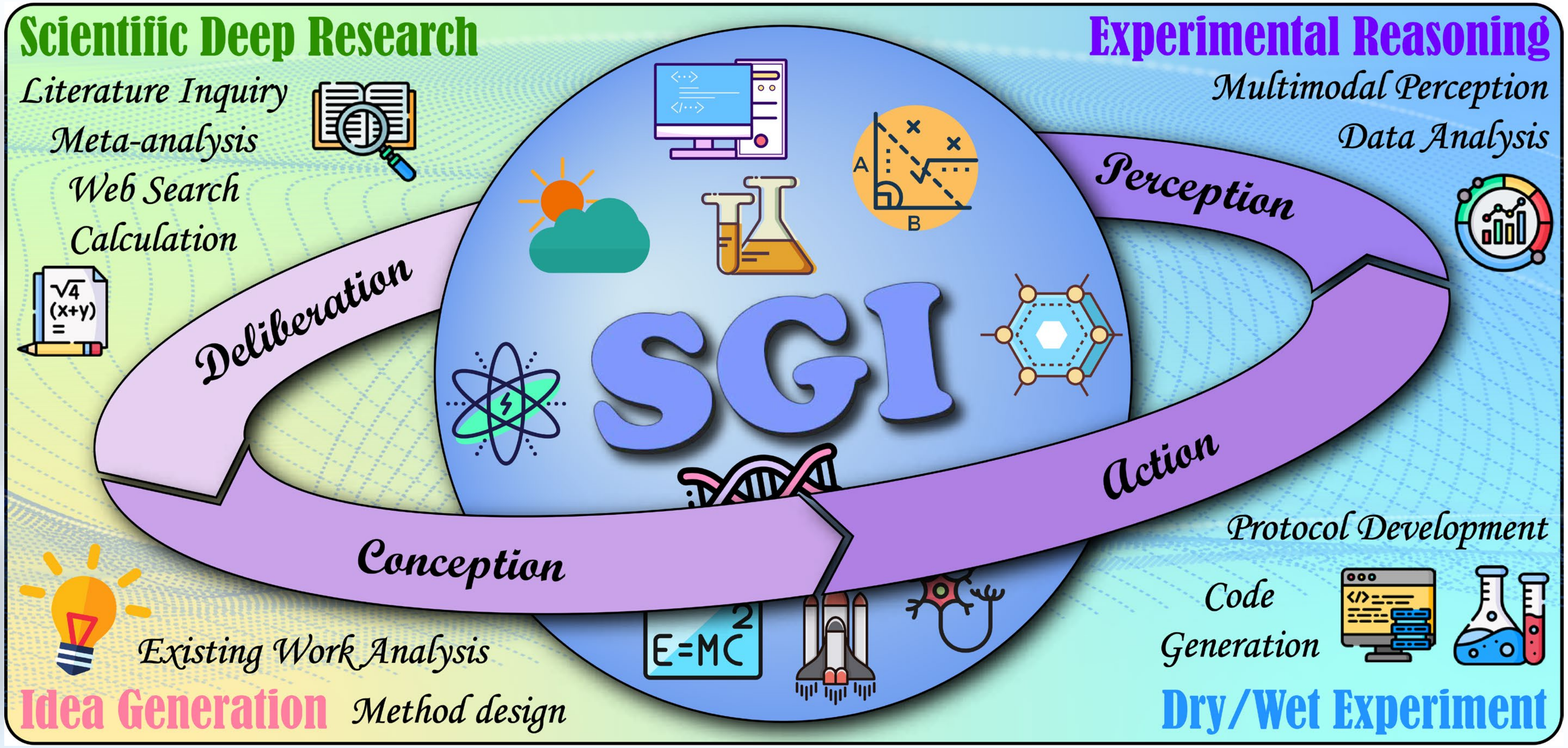
Action（行动）
写代码，做实验

定义SGI（通用科学智能）

SGI

能够自主完成科学探究的完整迭代周期

并展现出与人类科学家相媲美的灵活性和专业能力的人工智能



现有工作的局限性

单一学科 Benchmark

聚焦于某一特定的学科，例如化学（ChemBench）地球（EarthSE），物理（PHYSICS），海洋(OceanBench），数学（MATH，AIME-2025）。无法考察不同的学科方向。

单一任务 Benchmark

信息检索和整理任务（DeepResearch Bench），创意生成任务（MOOSE-Chem），工具使用任务（ToolUniverse），代码任务（SciCode），科学数据分析任务（SFE）。无法考察科研全流程。

评测AI的通用科学能力

- 需要多学科，多方向结合
- 需要多种不同的科研任务，能够与人类科学家的科研流程对齐

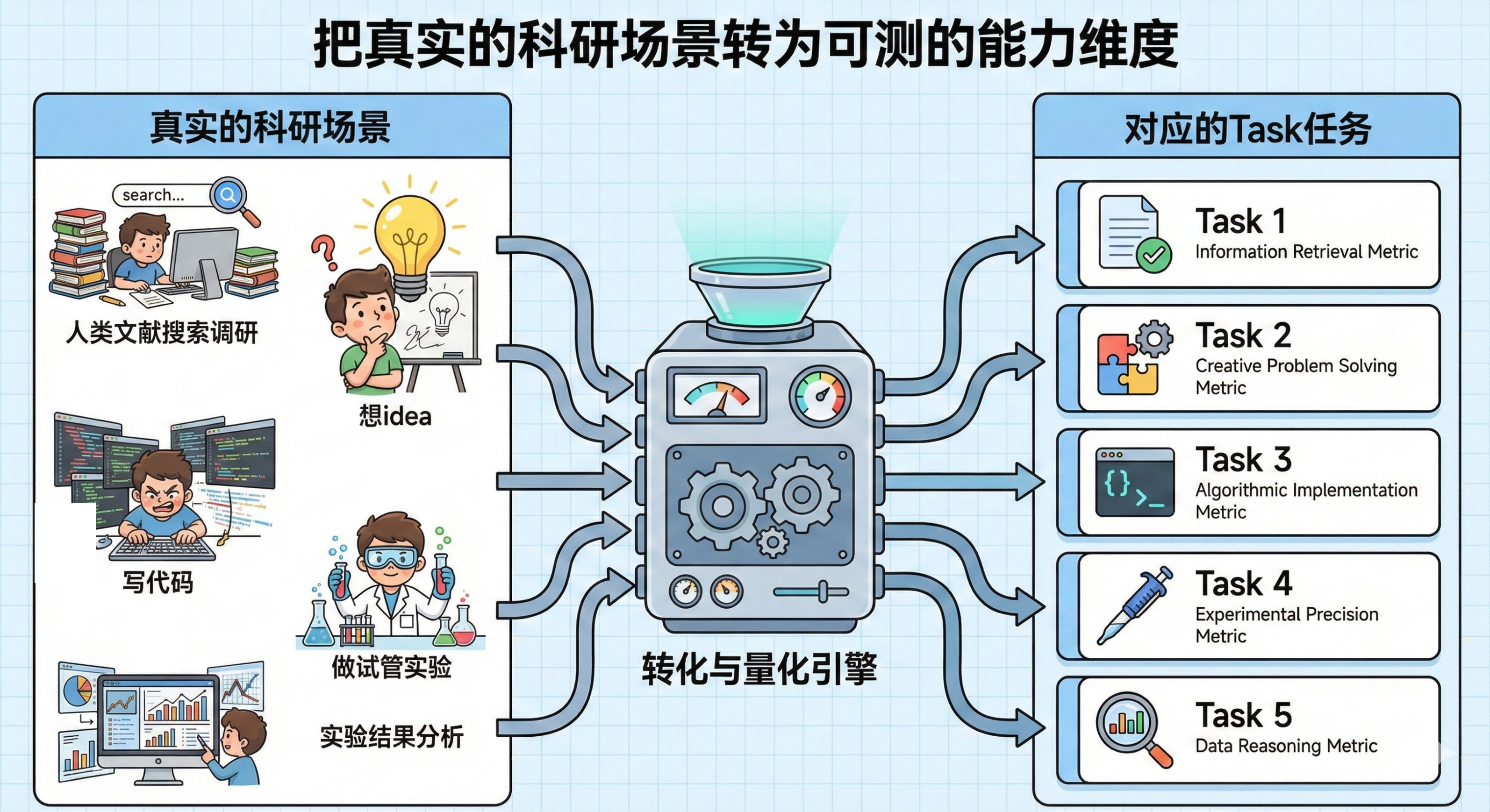
Benchmark构建

应用场景的真实性

规范任务的可测性

约束回答的开放性

确保题目的科学性



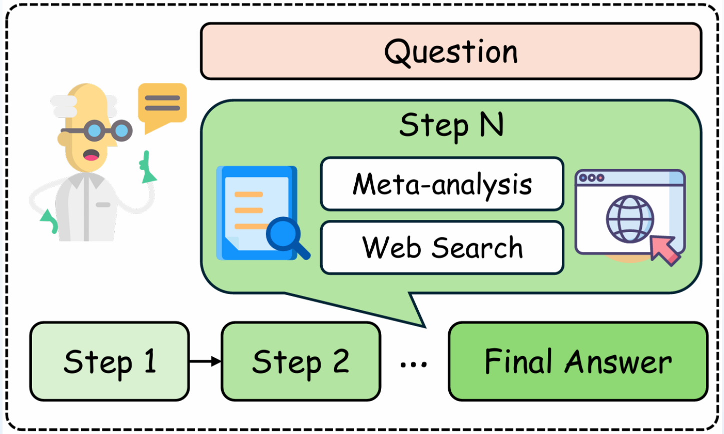
任务定义1：Deliberation——Scientific Deep Research

Task Input

· **背景**：研究主题的详细背景信息（学科、子领域等），用于避免术语歧义。
· **约束**：实验设定、科学假设、数据来源等限制条件。
· **数据**：任务中直接给出的实验或观测数据（可以是显式的，也可以是可推断的）。
· **问题**：需要回答的具体问题，例如某个量的数值或随时间的变化。
· **回答要求**：对答案形式的要求，比如单位、是否为整数、小数位数等。

Task Output

· **步骤**：系统为检索、处理数据或进行推理所给出的逐步解决方案。
· **答案**：精确的数值或字符串形式的最终回答（例如具体数值或短语）。



Question: The experimental methodology for studying chaotic hysteresis in Chua's circuit is employs a precision Chua's circuit setup with calibrated instrumentation to investigate chaotic hysteresis through step-by-step DC voltage variation and frequency-dependent triangular wave analysis, quantifying hysteresis loops and identifying critical frequency thresholds via phase space trajectory monitoring and time series bifurcation analysis. In the Chua circuit experiment, what are the calculated time constants (in μs) for the RC networks formed by a 10.2 nF capacitor C1 and the equivalent resistance, the peak-to-peak voltage (in V) range of the hysteresis loop at 0.01 Hz driving frequency, and the critical frequency (in Hz) where chaotic behavior ceases? Output the results in two decimal places, one decimal place, and integer format respectively, separated by commas.

User

Step 1: Find paper *Experimental observation of chaotic hysteresis in Chua's circuit driven by slow voltage forcing.*

Step 2: Identify RC network components from experimental setup: C1=10.2 nF, R1=219Ω. Calculate time constant: τ=R1×C1=219×10.2×10−9=2.2338μs≈2.23μs.

Step 3: Voltage range determination: At 0.01 Hz triangular forcing, peak-to-peak voltage ΔV_T=3.2 V measured from hysteresis loop width in experimental phase portraits.

Step 4: Critical frequency identification: "For f>10Hz the hysteresis phenomenon practically disappears" confirmed through frequency sweep experiments showing ΔV_T reduction from 3.2V (0.01Hz) to 0V (10Hz).

Step 5: Validate measurement procedures: Hysteresis loops are measured by "changing DC voltage very slowly and step by step" while monitoring attractor transitions between single scroll and double scroll regimes.

Step 6: Confirm data analysis techniques: Phase portraits and time series analysis confirm chaotic behavior through "bifurcations and dynamic attractor folding".

Step 7: Integrate experimental specifications: The setup uses DAQ CB-68LPR with LabVIEW, Sony AFG 320 function generator, and ±15V power supply for TL084CN op-amps, ensuring accurate voltage measurements.

Final Answer: 2.23, 3.2, 10

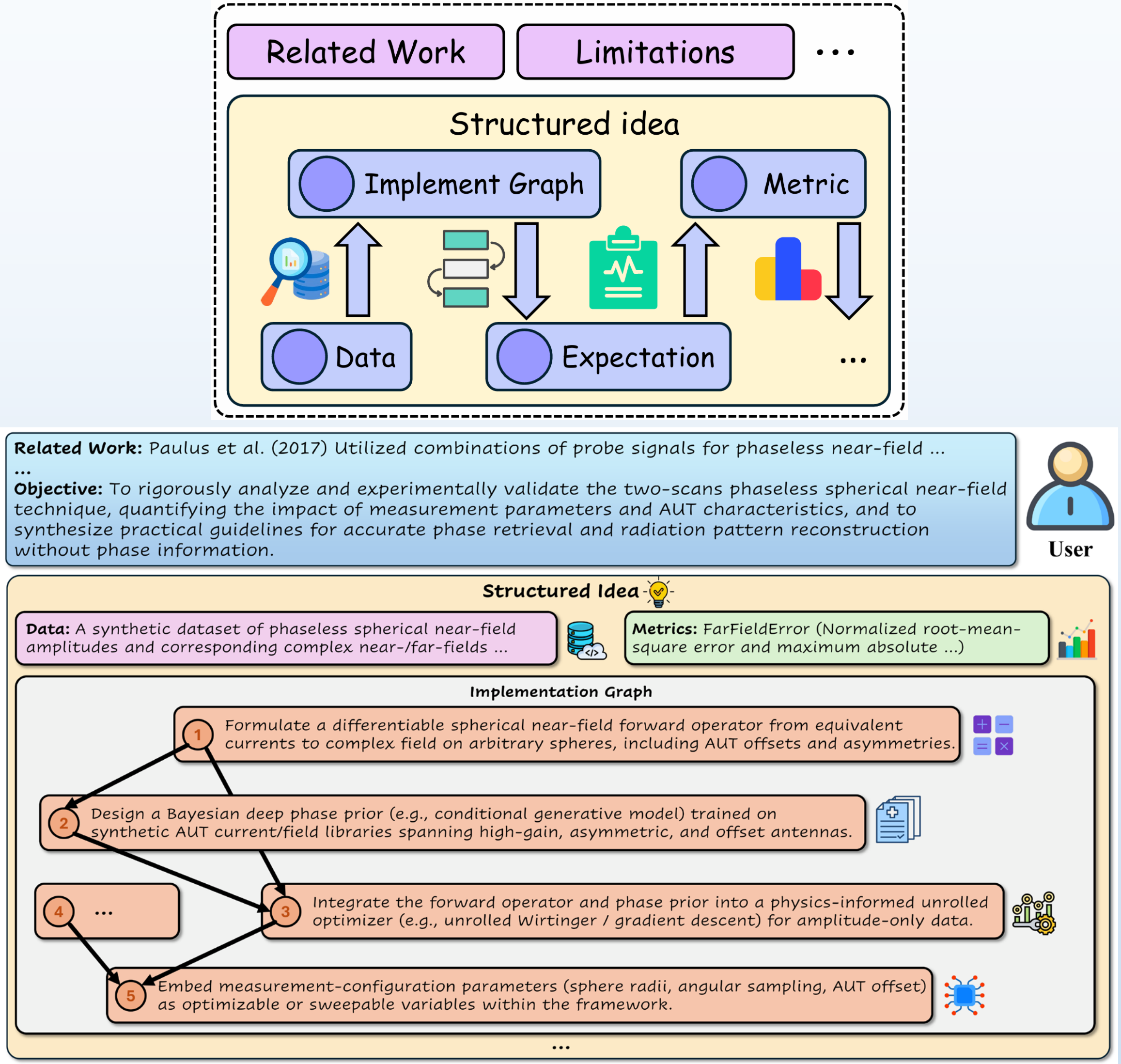
任务定义2：Conception——Idea Generation

Task Input

· **相关工作：**与该研究方向相关的已有研究总结，为新想法提供上下文。
· **挑战：**当前领域面临的难题以及现有方法的不足之处。
· **局限：**现有研究的具体限制或缺陷，新想法需要针对这些问题。
· **动机：**从何种视角、出于何种原因希望解决上述局限。
· **任务目标：**本任务的主要目标，例如生成能解决挑战或改进现有方法的想法。
· **现有解决方案：**目前该领域常用的方案或方法描述。

Task Output

· **核心想法：**用来解决研究挑战的中心新思想或新概念。
· **实现步骤：**实现该核心想法所需的关键步骤或流程。
· **实现顺序：**上述步骤的执行先后顺序。
· **数据：**实现和评估该想法所需使用的数据。
· **评价指标：**用于衡量该想法效果或价值的评估标准。
· **预期结果：**预计该方法或想法能取得的成果或贡献。



任务定义3.1：Action——Dry Experiment

Task Input

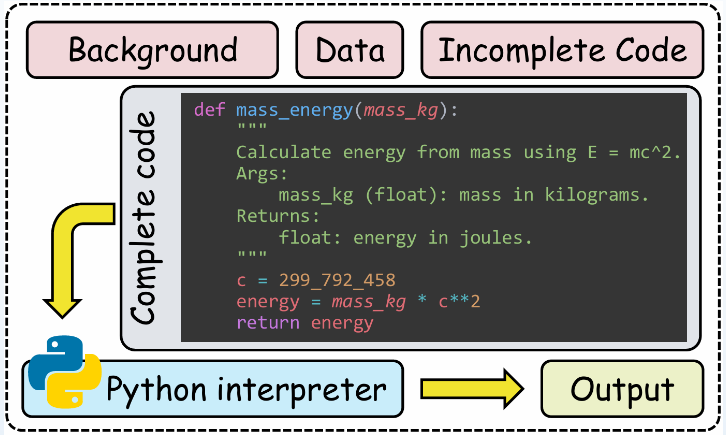
· **背景：** 相关科学代码及问题背景信息，为于实验提供上下文。

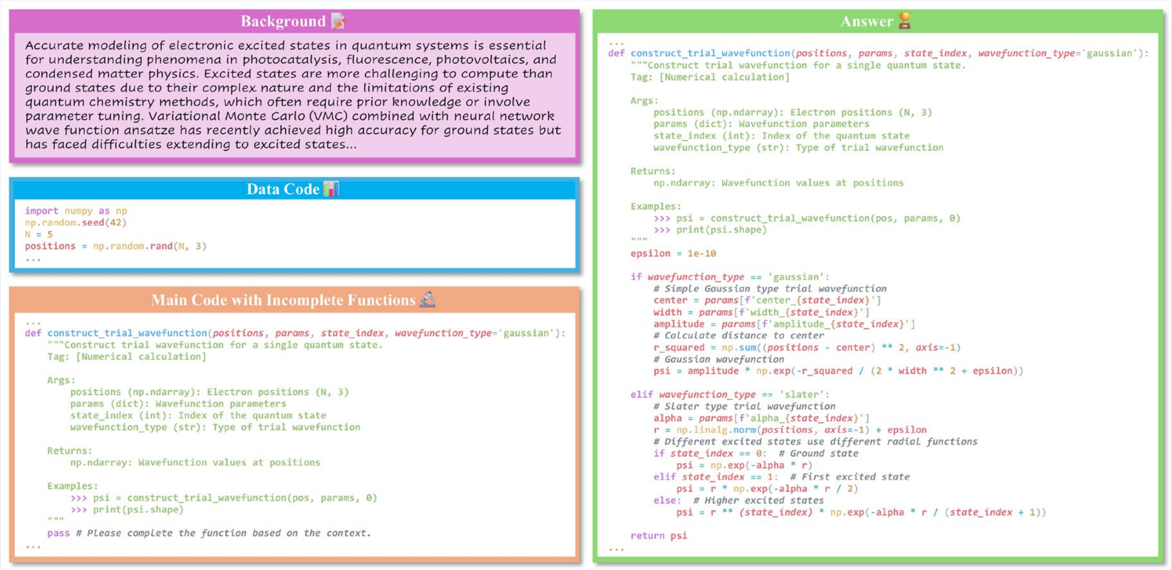
· **数据代码：** 实验中使用的数据及其对应代码片段或预定义输入。

· **主体代码：** 实验的主代码框架，其中部分函数被遮蔽或缺失。

Task Output

· **函数：** 需要由模型补全或生成的缺失函数，实现主体代码中的关键逻辑。





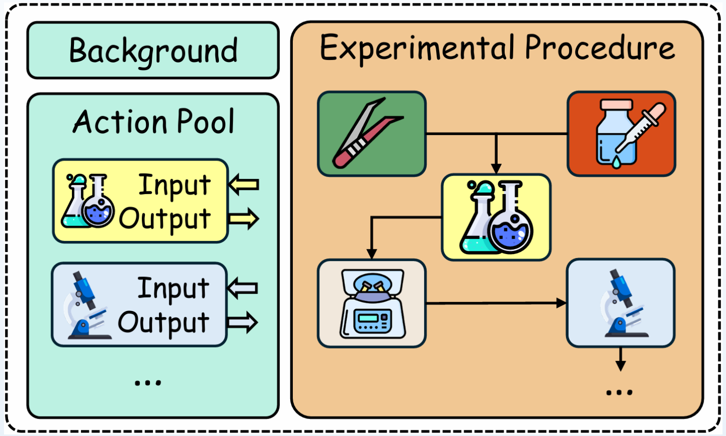
任务定义3.2：Action——Wet Experiment

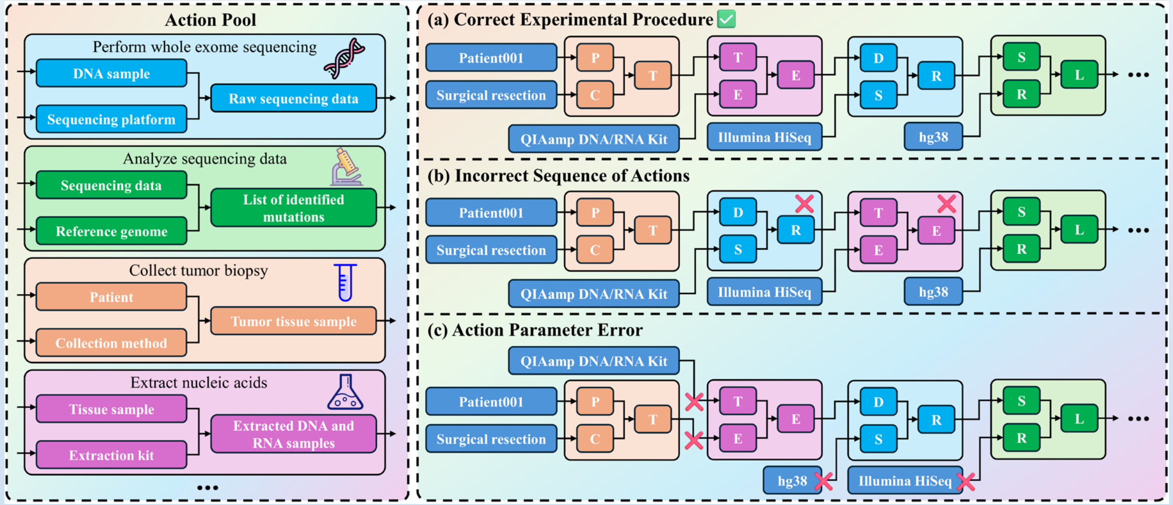
Task Input

- **背景：**相关实验流程或操作背景说明。
- **动作池：**可用于实验的一组原子动作，包含每个动作的说明及其输入/输出定义。

Task Output

- **原子动作顺序：**各原子动作在实验中应执行的顺序。
- **原子动作参数：**每个原子动作对应的参数设置（例如试剂种类、浓度、体积、温度等）。





任务定义4：Perception——Experimental Reasoning

Task Input

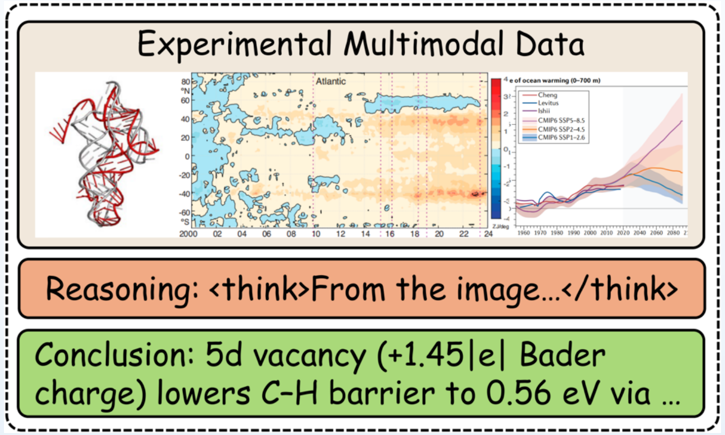
· **多张实验图像：**表示实验结果或仪器观测数据的一组图像（可包含流程图、观测图、实验装置图、仿真图、可视化图等）。

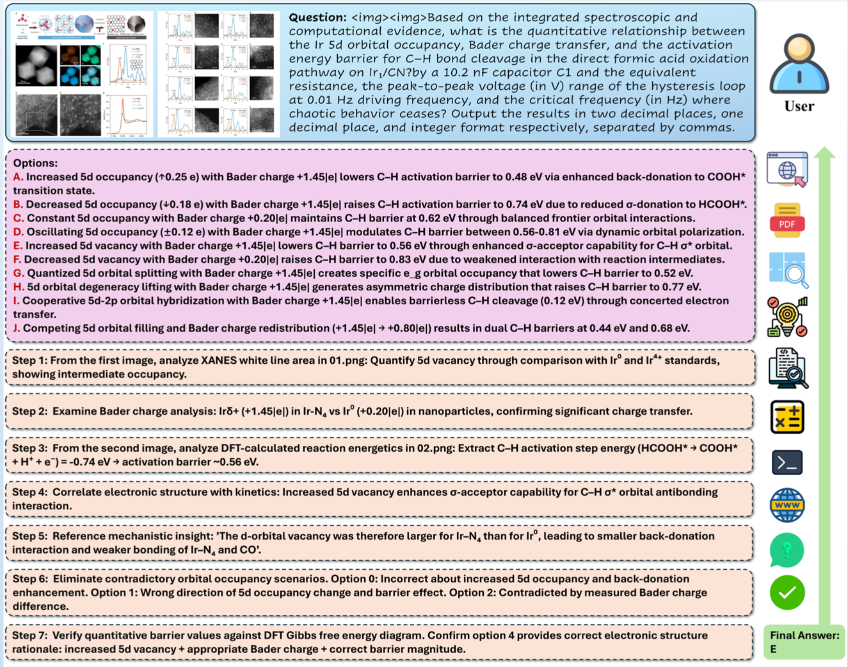
· **问题：**与这些实验数据/图像相关的具体问题或待验证假设。

Task Output

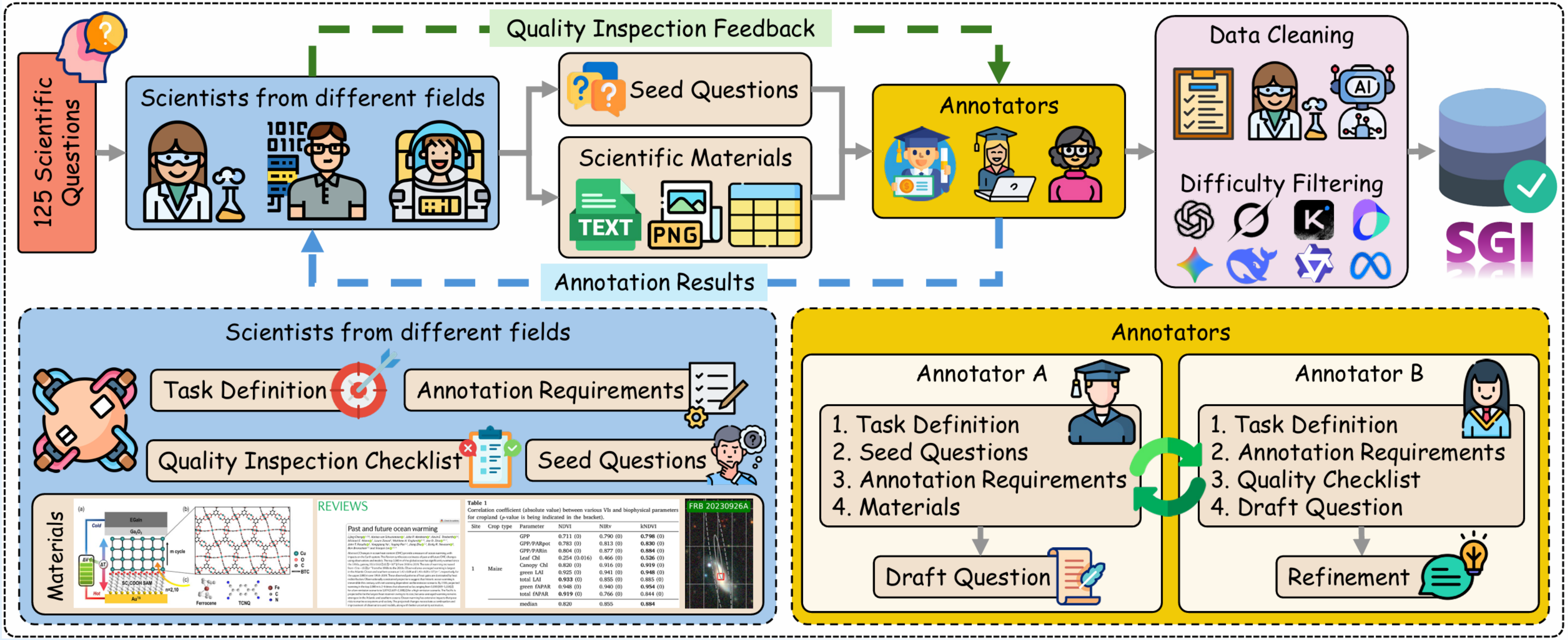
· **推理过程：**为得到答案而进行的逐步推理，包括计算、分析、比较等。

· **答案：**在综合多模态证据后给出的最终结论，用于回答给定问题或假设。





数据构建：Scientist-Aligned Data Construction



1. 领域专家提供**原始素材，种子问题**，制定构建规则
2. 硕士或**博士生**根据问题规则进行问题构建
3. 数据清洗：**规则校验，模型校验，专家复核**
4. 难度筛选：**模型测试**，除过于简单的题目

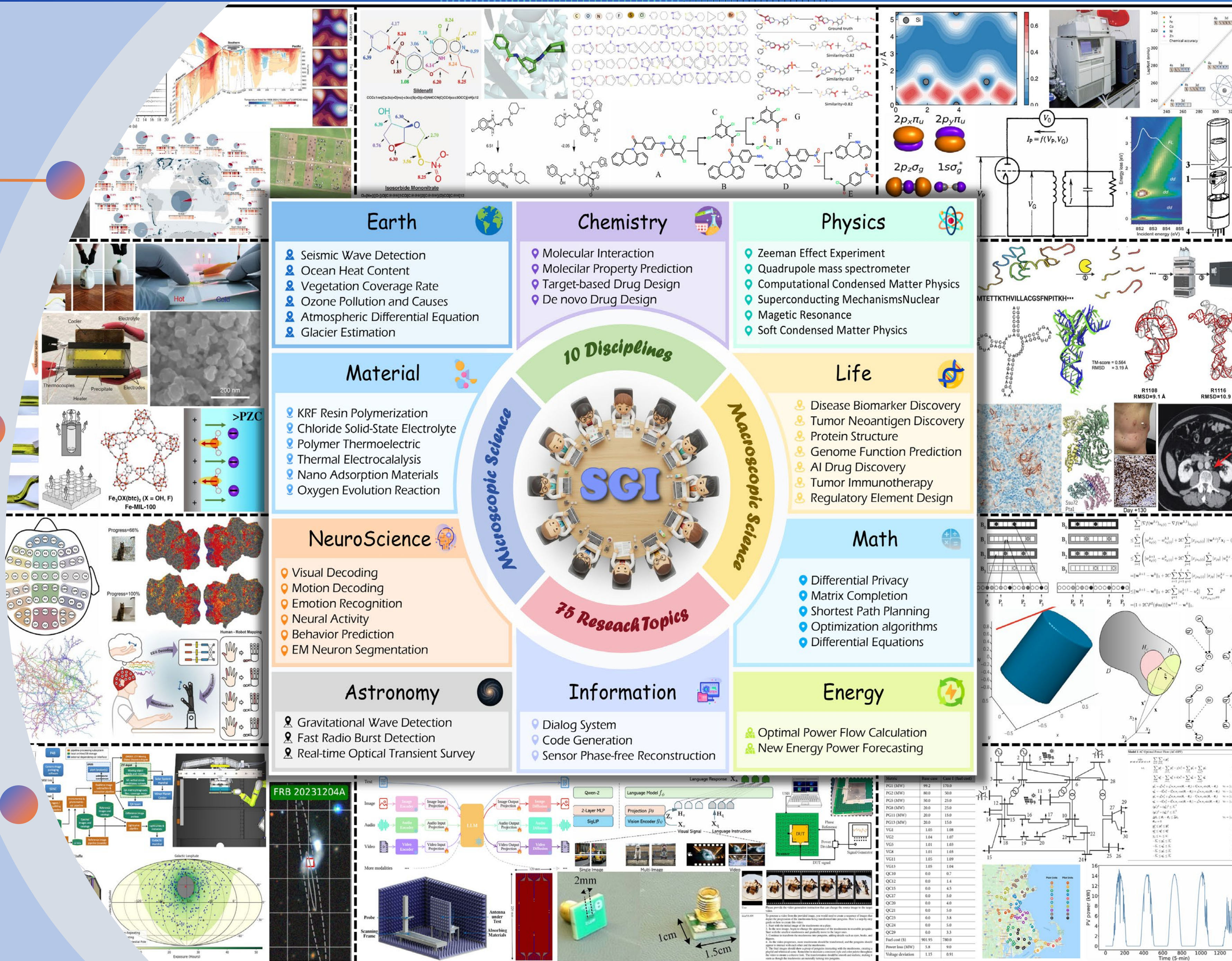
学科分布

10大学科：

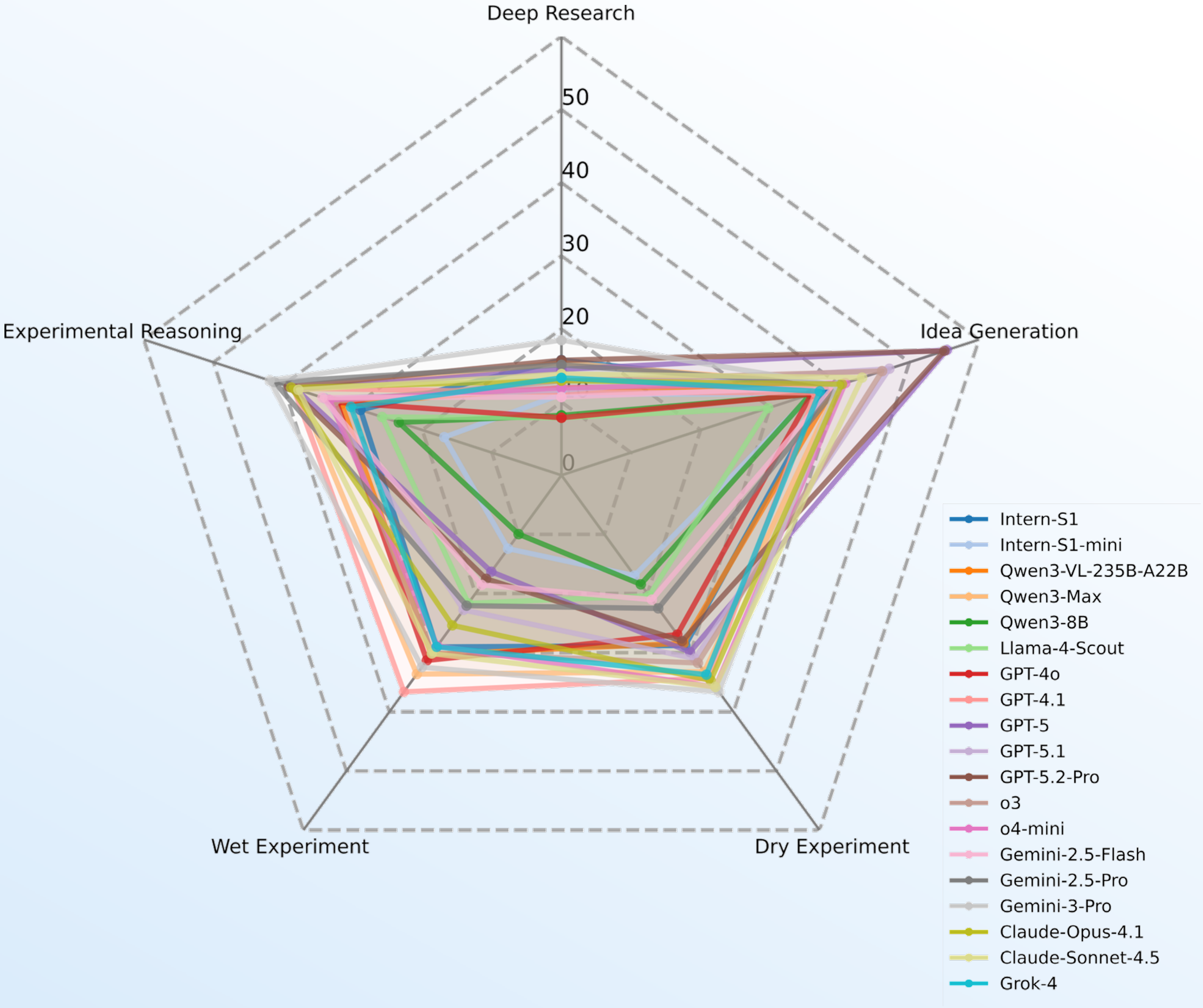
地球，化学，物理，生命，数学，能源，信息，天文，神经科学，材料

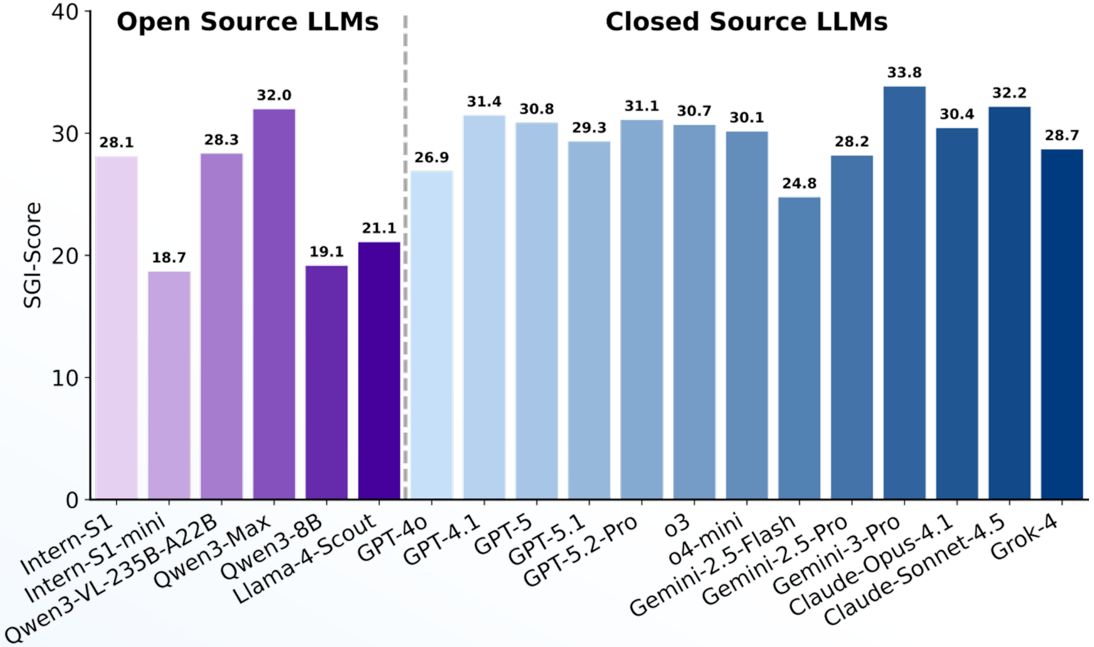
75个研究方向：

海洋热含量，药物设计，塞曼效应，蛋白质结构预测，差分隐私，能源概率预测，传感器设计，引力波检测，视觉解码，树脂聚合反应，…



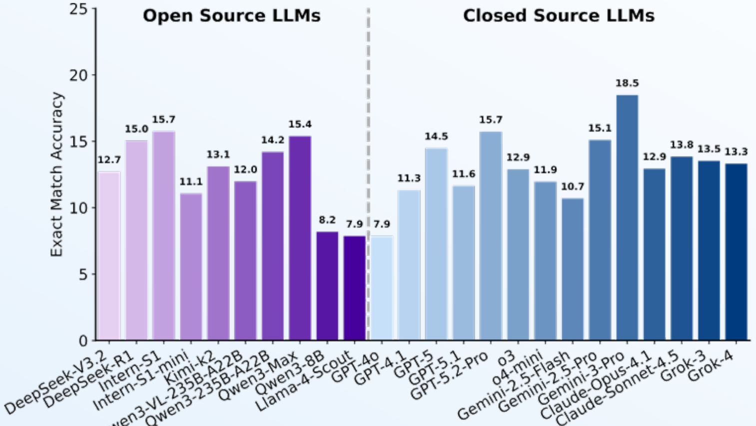
评测结果总览

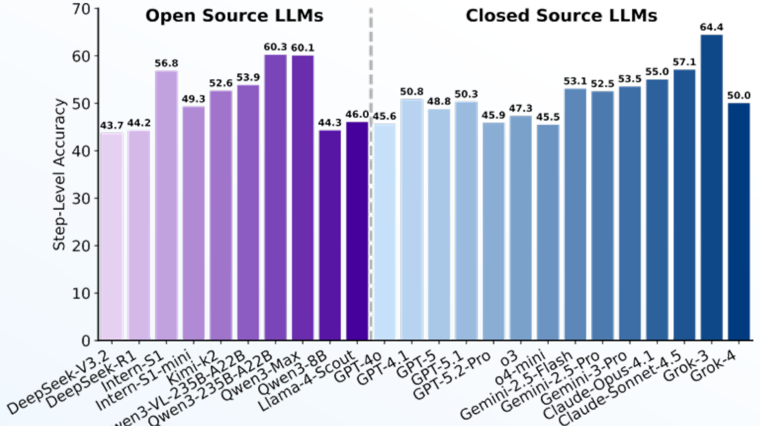


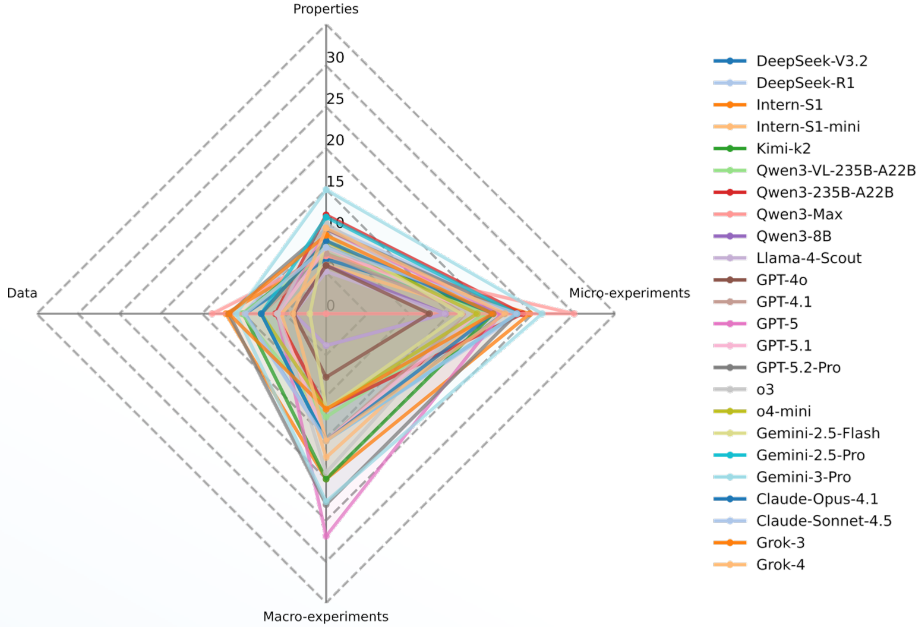


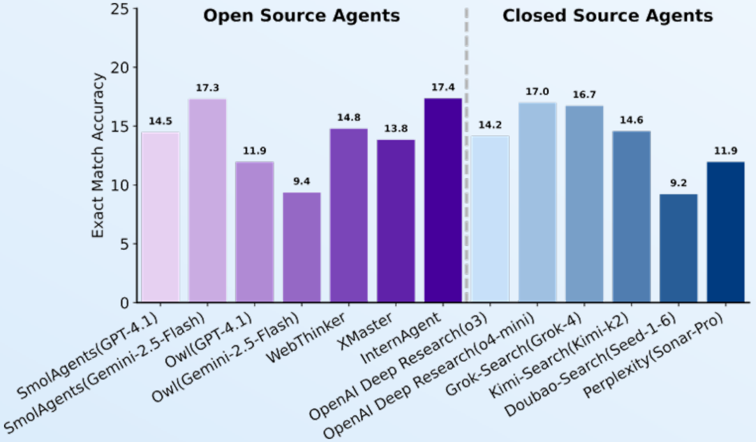
1. **Deep Research：**长链路推理易崩塌；
2. **Idea Generation：**新颖度尚可，但可行性弱；
3. **Dry Experiment：**能跑通但算不对；
4. **Wet Experiment：**时序与分支协调困难；
5. **Experimental Reasoning：**对比型推理最难。

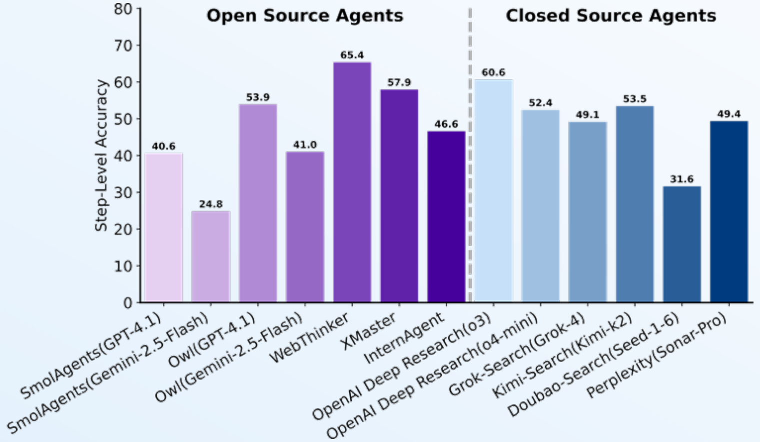
任务1： Scientific Deep Research











1. 步骤准确率达 50%–65%，但**长链条步骤中的错误导致最终结论频繁错误**，答案严格匹配仅 10%–20%；
2. 工具增强的多智能体在逐步准确率略优，但与纯模型**差距并不显著**；
3. 类型上，**“数据/性质”题最难**，需跨文献精确检索与数值聚合；“实验”类相对较好但整体仍低于 30%，体现元分析的严苛性。

任务1： Scientific Deep Research

题目

基于IAP/CAS 数据集对全球 0–2000 m 海洋热含量（OHC）的分析：
1958-2021年的平均年增暖率记为R_1
1986-2021年的平均年增暖率记为R_2
问：1986–2021 时期的年增暖率，相比 1958–2021 时期增加了百分之多少？

OpenAI Deep Research o4-mini 的回答

1. 模型查到另一份 IAP/CAS OHC 资料，给出 **1955–2023** 的趋势约 6.4 ZJ/yr。❌
2. 模型对 1958–2021 做线性拟合：R_1 ≈6.41 ZJ/yr ❌
3. 模型对 1986–2021 做线性拟合：R_2 ≈7.81 ZJ/yr ❌
4. 计算相对增加百分比：$\left(\frac{7.81}{6.41}-1\right)\times 100\% \approx 21.88\%$ ❌

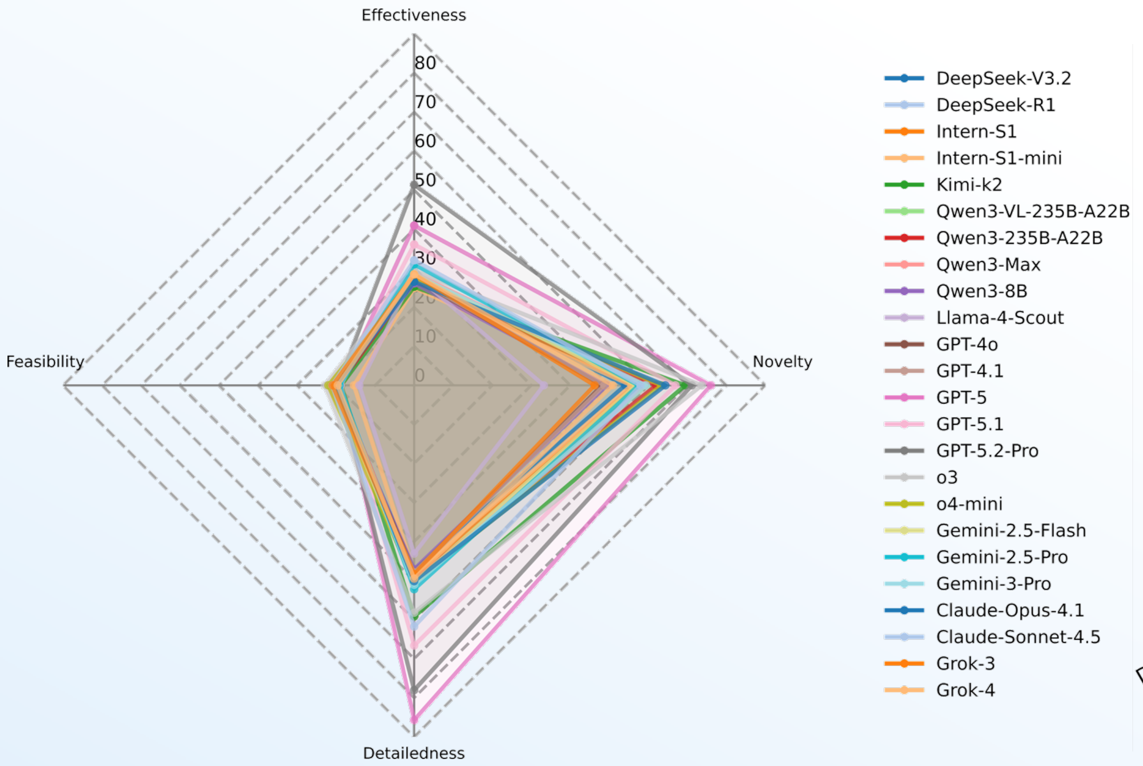
模型使用的资料的时间区间与题目要求不一致

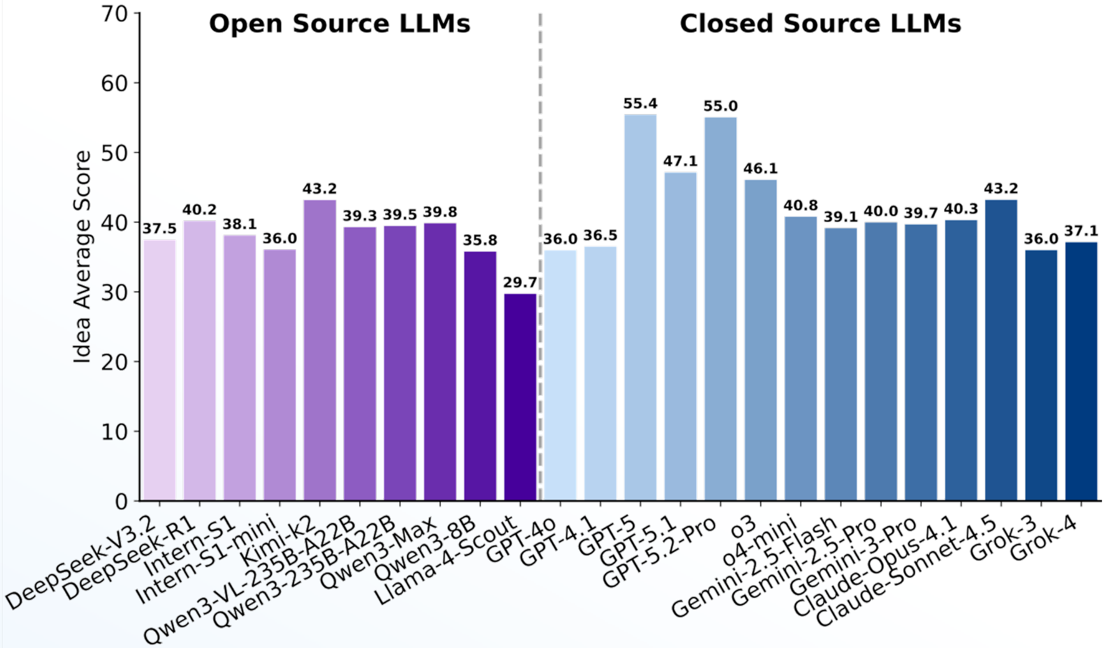
错误原因

1. Ocean Warming Continues through 2021 despite La Niña Conditions
2. R_1 = 5.7 ZJ/yr（1958–2021）
3. R_2 = 9.1 ZJ/yr（1986–2021）
4. 计算相对增加百分比
$\left(\frac{9.1}{5.7}-1\right)\times 100\% \approx 59.65\%$ ✅

正确答案

任务2：Idea Generation





1. 闭源模型“新颖性（Novelty）”更强，但“可行性（Feasibility）”普遍偏低。以 GPT-5 为例：新颖性 76.08、可行性 18.87，体现“**概念丰富 ≠ 可执行方案**”；
2. 开源可行性上限约 20 分（如 Qwen3-Max 20.98），多数模型 14–20 分，显示**“能说清”与“能落地”之间的落差**；
3. 常见缺陷：**缺少数据获取与预处理计划**；流程接口不闭合（输入输出不对齐）；**步骤顺序与依赖模糊**，导致“创意→蓝图→执行”闭环断裂。

任务2：Idea Generation

研究目标

定量比较 RAFT 与 FRP 制备的光刻胶共聚物三元共聚物的组成非均一性

分析重点

- 单体序列分布（monomer sequence distribution）
- 组成漂移（compositional drift）
- 组成均匀性对光刻胶性能（尤其 LER）的影响

人类专家的Idea

含具体的参数和方法名称，前后逻辑紧密✅

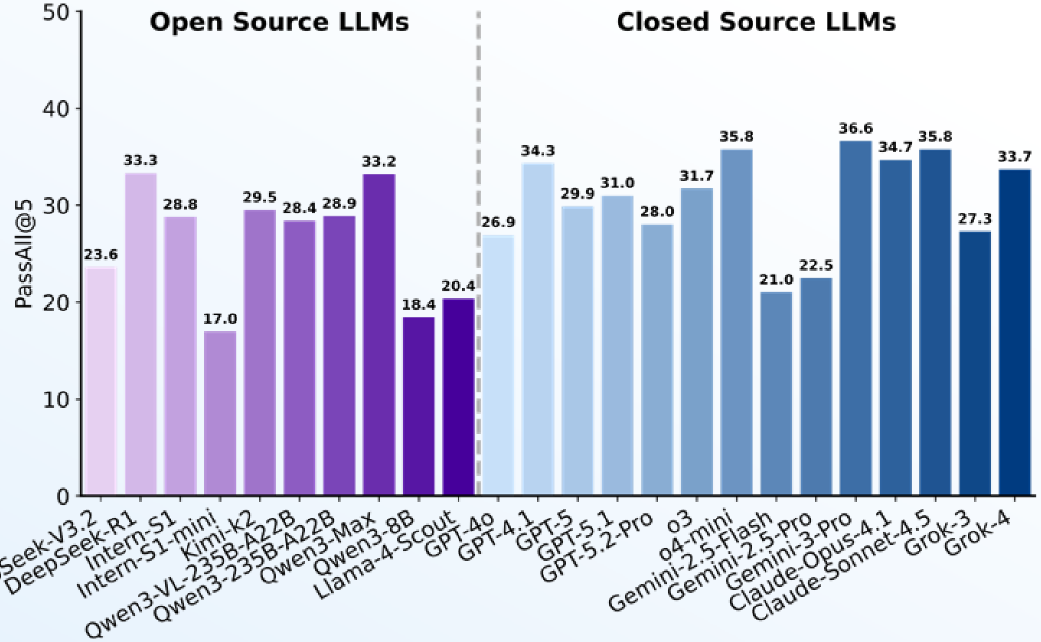
- 对比实验与理论组成并评估随转化率的漂移
  - 准备并纯化 AOST/Sty/tBA 单体与 RAFT 试剂 EM左边层级
  - 在 65°C 下用 FRP 与 RAFT 合成共聚物/三元共聚物并变化进料比左边层级
  - 左低转化率（<10%）淬灭来避免组成漂移边层级
  - 左通过沉淀和干燥纯化聚合物产物边层级
- 分析组成与序列均匀性对光刻胶 LER 的影响
  - 用 $^1H/^{13}C$ NMR 对聚合物组成进行定量分析
  - 用 SEC 测 Mn 与 Đ 比较 RAFT/FRP 控制能力。
  - 用 NLLS（REACT/CONTOUR）拟合共聚反应比
  - 用 Alfrey–Goldfinger 与概率模型预测序列分布

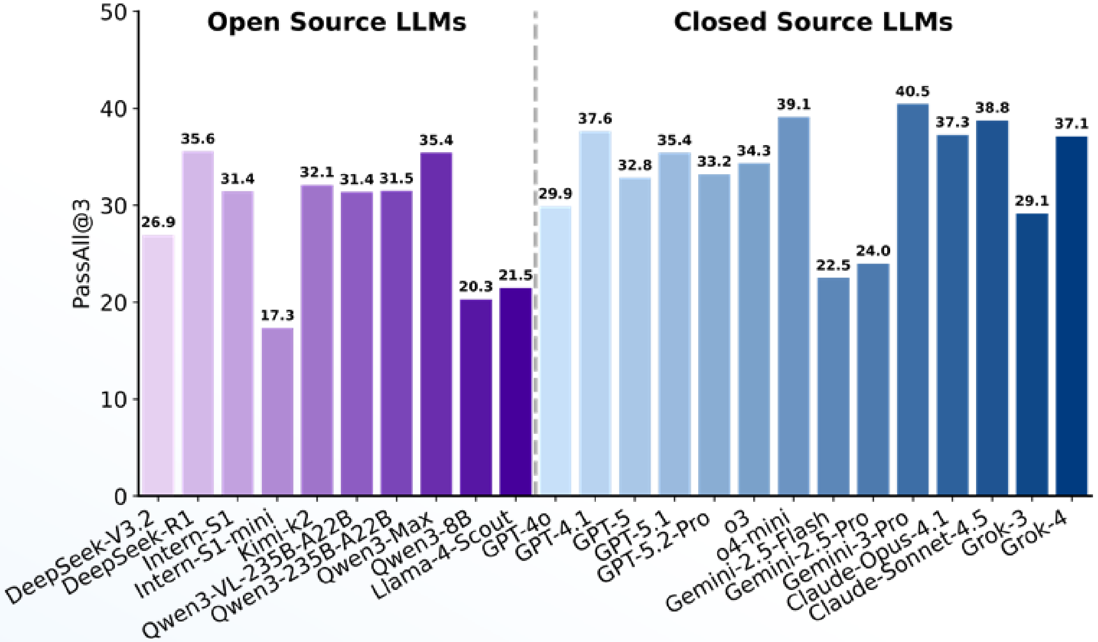
GPT-5.1的Idea

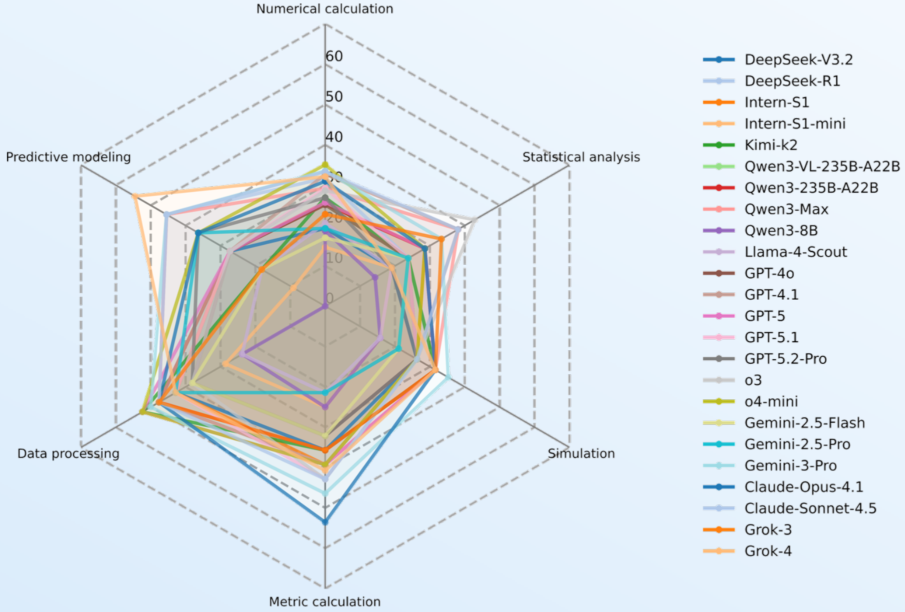
缺乏具体工作流，具体参数，具体方法❌

- 探索序列结构与溶解行为的关联
  - 选择一组可能适用于光刻胶的模型单体体系并初步设定目标结构
  - 尝试通过多种分析手段获取精细的反应比数据（NMR / UPLC 等
- 测量 LER/LWR 等关键图形参数
  - 构建一个 kMC + RAFT 平衡序列预测模型
  - 基于模型推断设计时间变化进料以获得多种理想化序列结构
  - 在反应过程中进行原位监测（NIR/Raman）
- 建立序列特征到 LER 的预测模型
  - 通过高级 NMR、MS、数据解卷积等方法解析序列分布
- 基于模型进行序列优化并验证
  - 在树脂体系中控制 PAG 分布并利用 TOF-SIMS / STEM-EDS 进行空间分析
  - 在 EUV/193i 全流程中测试材料性能

任务3.1：Dry Experiment







1. 每题含 5 个单测，最佳 Gemini-3-Pro 的严格通过率（全过 5 个单测）仅 36.64%，宽松通过率（至少过 1 个）41.98%，表明**模型常能写对部分，但难以实现严格正确**；
2. 闭源模型略优于开源，但**优势有限且分布重叠**，“科学代码合成”仍是各架构共同短板；
3. 平滑执行率（无报错运行）多在 90%+，显示“能跑”与“算对”之间存在鸿沟；
4.类型上，数据处理/预测建模较稳；数值计算与仿真最弱，受离散化、稳定性与约束处理影响。

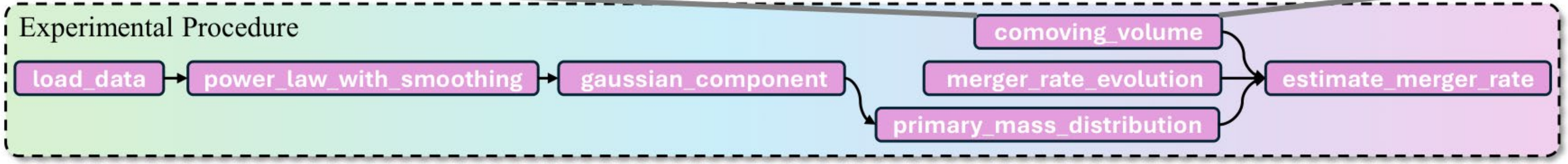
任务3.1：Dry Experiment

GPT 4.1 Response

```python
...
def comoving_volume(z_max: float, H0_val: float, omega_m_val: float, omega_l_val: float) -> float:
    """
    Calculates the comoving volume up to a maximum redshift z_max.
    Tag: [Numerical calculation]

    Args:
        z_max (float): The maximum redshift to integrate to.
        H0_val (float): The Hubble constant in km/s/Mpc.
        omega_m_val (float): The matter density parameter.
        omega_l_val (float): The dark energy density parameter.

    Returns:
        float: The comoving volume in Gpc^3.
    """
    # Define the integrand for comoving distance integral
    def E(z):
        return np.sqrt(omega_m_val * (1 + z)**3 + omega_l_val)

    def integrand(z):
        return 1.0 / E(z)

    # Speed of light in km/s
    c_km_s = C

    # Integrate to get comoving distance in Mpc
    integral, _ = integrate.quad(integrand, 0, z_max, epsabs=1e-8, epsrel=1e-8) ✅High precision

    d_c = (c_km_s / H0_val) * integral  # in Mpc

    # Comoving volume in Mpc^3 (assuming flat universe)
    volume_mpc3 = (4.0 / 3.0) * np.pi * d_c**3

    # Convert to Gpc^3
    volume_gpc3 = volume_mpc3 / 1e9

    return volume_gpc3
...
```

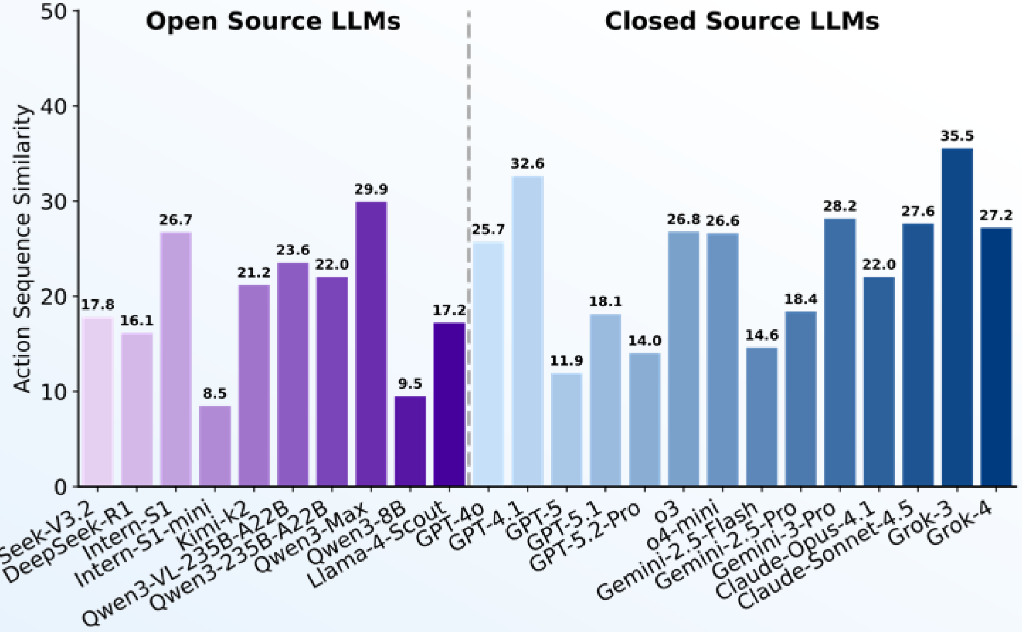
O4-mini Response

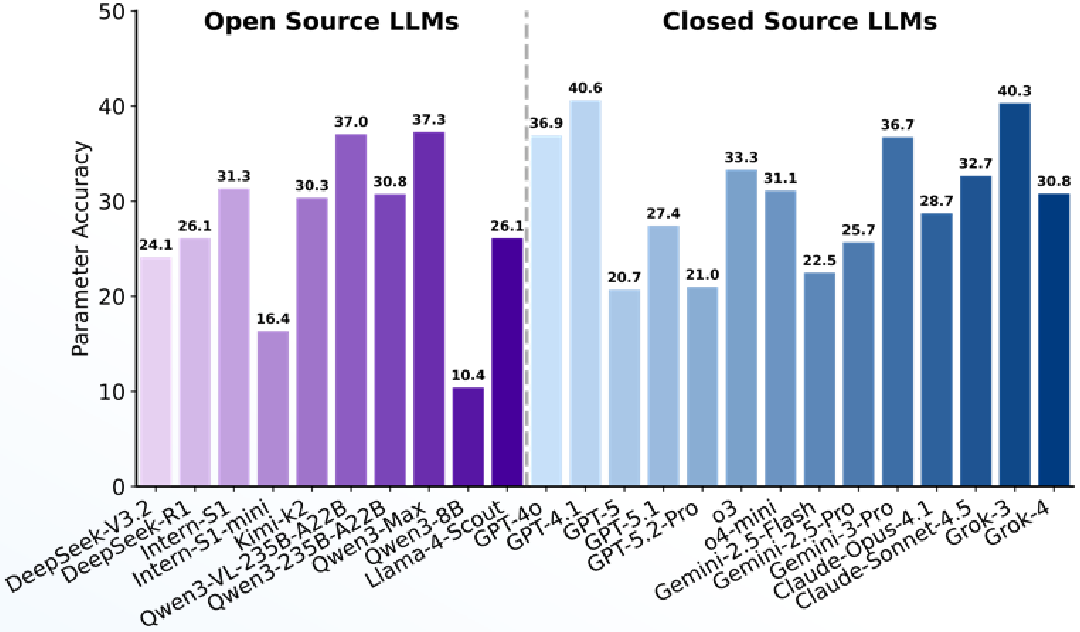
```python
...
def comoving_volume(z_max: float, H0_val: float, omega_m_val: float, omega_l_val: float) -> float:
    """
    Calculates the comoving volume up to a maximum redshift z_max.
    Tag: [Numerical calculation]

    Args:
        z_max (float): The maximum redshift to integrate to.
        H0_val (float): The Hubble constant in km/s/Mpc.
        omega_m_val (float): The matter density parameter.
        omega_l_val (float): The dark energy density parameter.

    Returns:
        float: The comoving volume in Gpc^3.
    """
    # Set up redshift grid
    nz = 2000
    z = np.linspace(0.0, z_max, nz)
    dz = z[1] - z[0]

    # E(z) = sqrt(Ω_M (1+z)^3 + Ω_L) for flat universe
    E = np.sqrt(omega_m_val * (1.0 + z)**3 + omega_l_val)
    # Comoving distance element dχ/dz = c / H0 / E(z) in Mpc
    dchi_dz = (C / H0_val) / E  # C [km/s], H0 [km/s/Mpc] ⇒ result in Mpc

    # Cumulative comoving distance χ(z) via simple trapezoidal integration
    chi = np.cumsum(dchi_dz) * dz ❌Low precision
    # Ensure χ(0)=0
    chi[0] = 0.0
    # Differential comoving volume element dV/dz = 4π χ(z)^2 * dχ/dz
    dV_dz = 4.0 * np.pi * chi**2 * dchi_dz  # in Mpc^3 per unit z
    # Integrate over redshift to get volume
    V_mpc3 = np.trapz(dV_dz, z)  # in Mpc^3
    # Convert to Gpc^3
    V_gpc3 = V_mpc3 / 1e9

    return V_gpc3
...
```

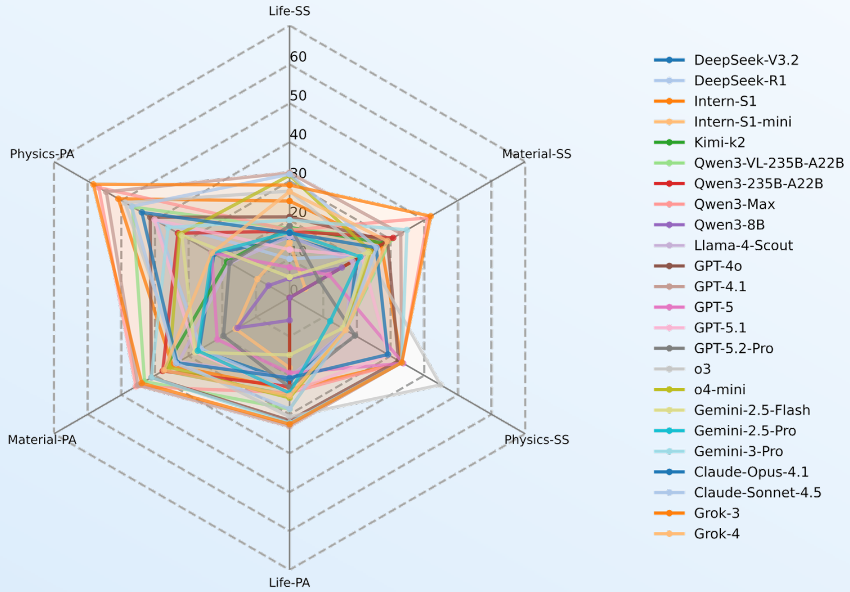


引力波体积估计中，前向累加（np.cumsum）与自适应积分（scipy.integrate.quad）差异巨大；**前者累积误差经 χ(z) 影响 dV/dz，最终体积严重偏离**。

任务3.2：Wet Experiment

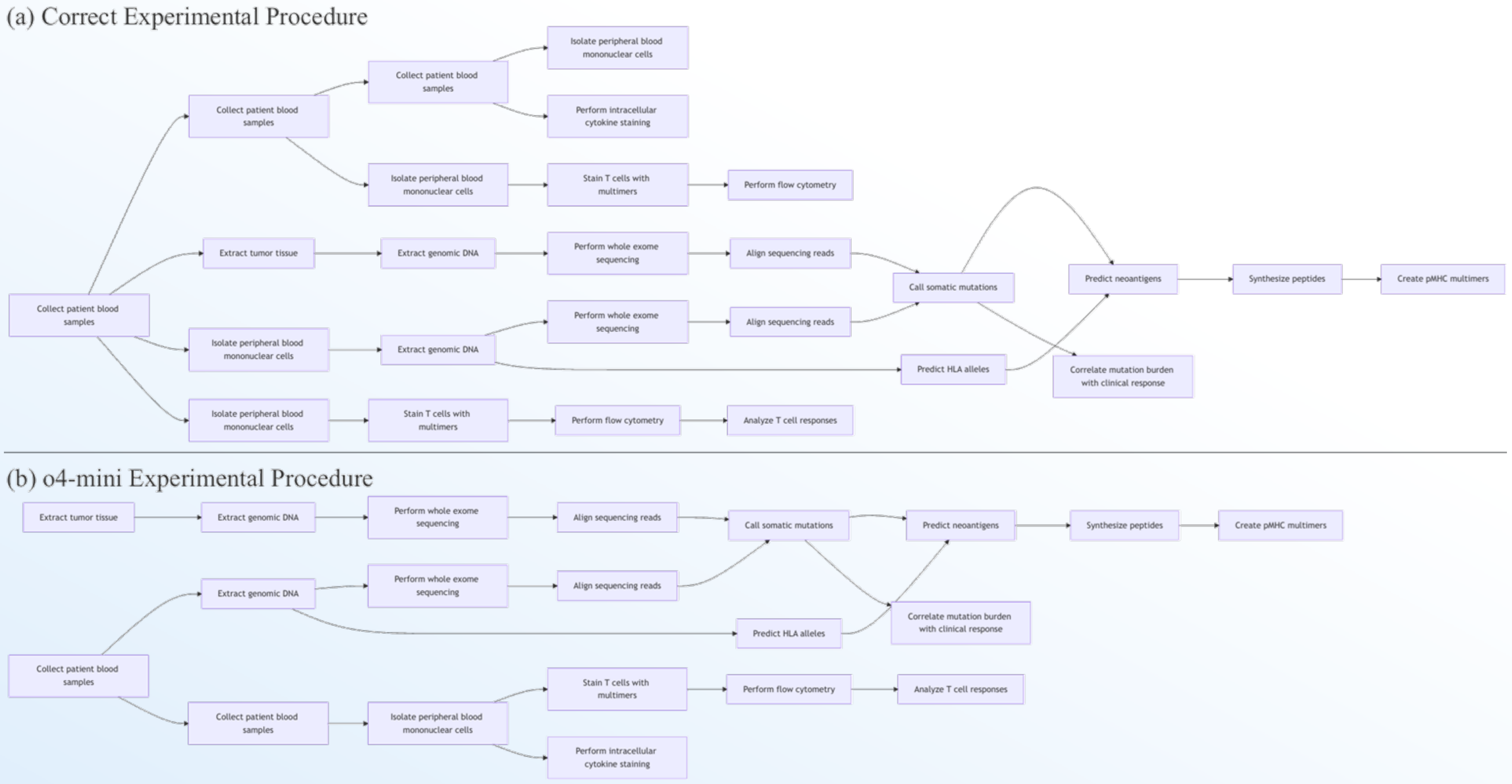






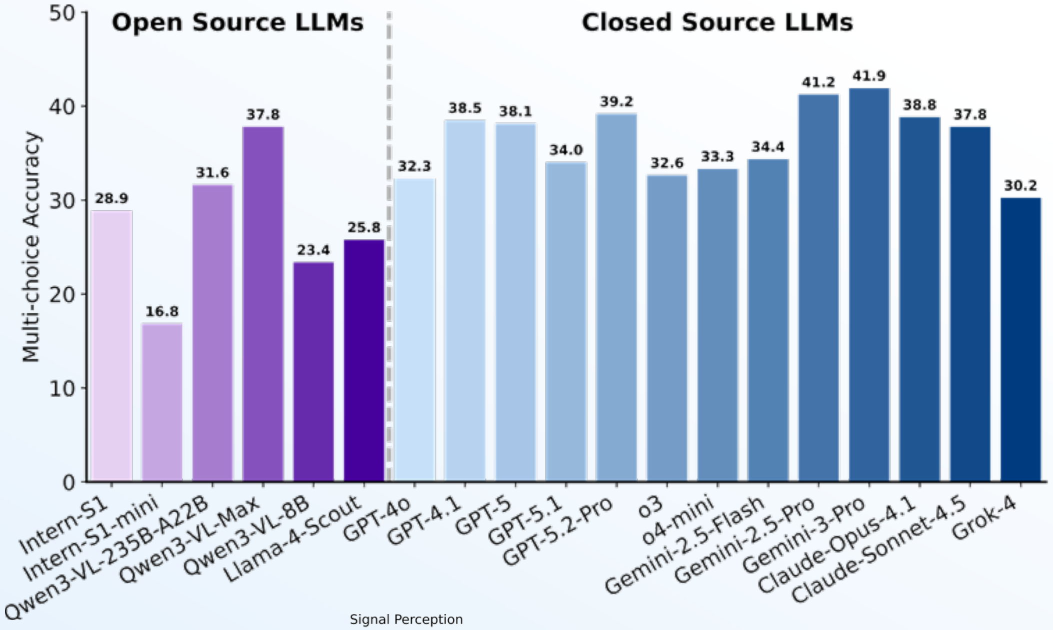
1. 序列相似度整体偏低，最佳闭源约 35.5；参数准确率最高约 40.6；部分闭源参数准确率显著下跌（约 20.7）；
2. 高发错误：**插入多余步骤、遗漏关键步骤、打乱有效步骤顺序**。

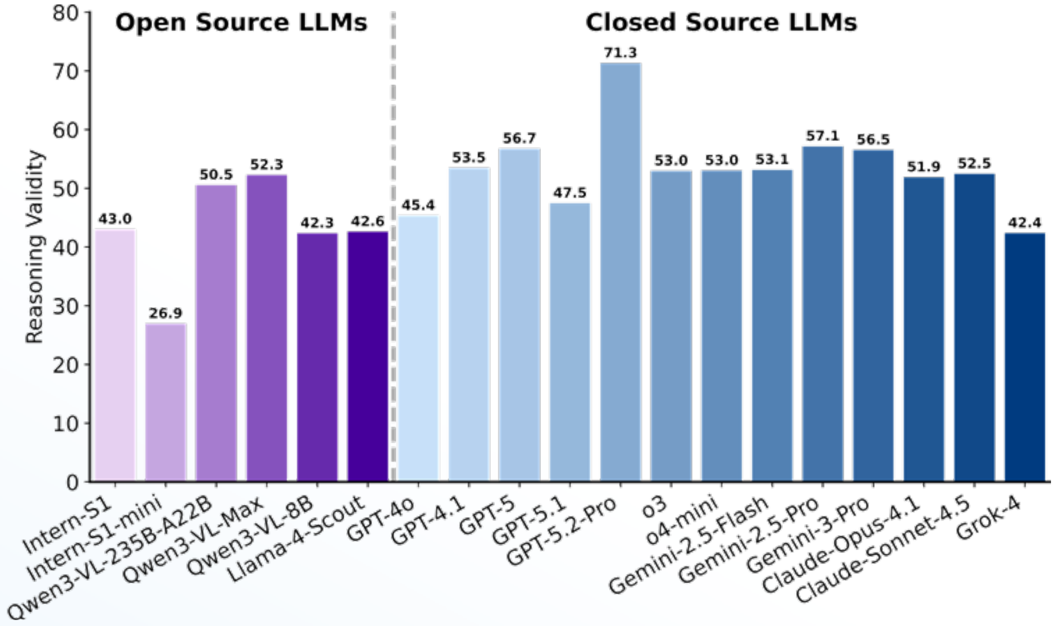
任务3.2：Wet Experiment

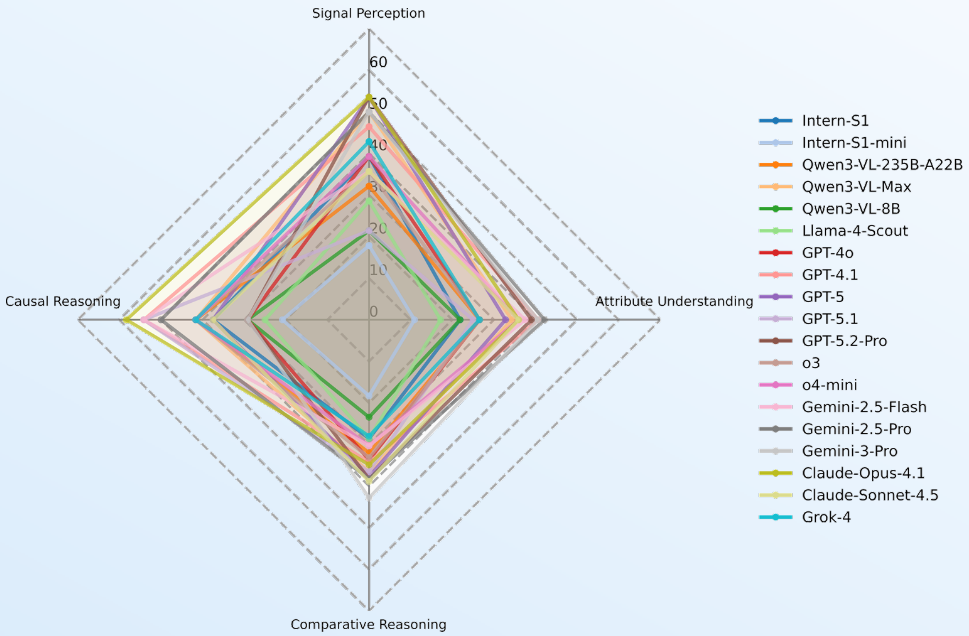


在 NSCLC 抗 PD-1 流程中，常见错误包括：将纵向采样简化为一次采血；PBMC 只在单一时间点分离；功能测定未按时间/刺激分组；**基因组测序与免疫表型流程混用样本**等，反映**时间协调、分支规划与样本管理**薄弱。

任务4：Experimental Reasoning







1. 闭源整体更强：最佳闭源答案准确率约 41.9、推理有效性最高约 71.3；
2. **多数模型推理有效性高于答案准确率**：难以实现推理链条的完全正确；
3.推理类型上，因果推断与感知识别较稳；**比较型最弱**，涉及跨样本细粒度对比与一致性判别。

任务4：Experimental Reasoning

问题 ❯❯❯

在主震震中西北方向（沿剖面距离约 15–25 km）处，大的同震滑动发生在浅部（约 0–5 km），理论上会导致该区域浅部余震数量偏少（余震亏损）。在这个区域里，哪一个余震目录最符合“**浅部大滑动 → 浅部余震偏少**”的预期模式？

Claude-Opus-4.1推理步骤 ❯❯❯

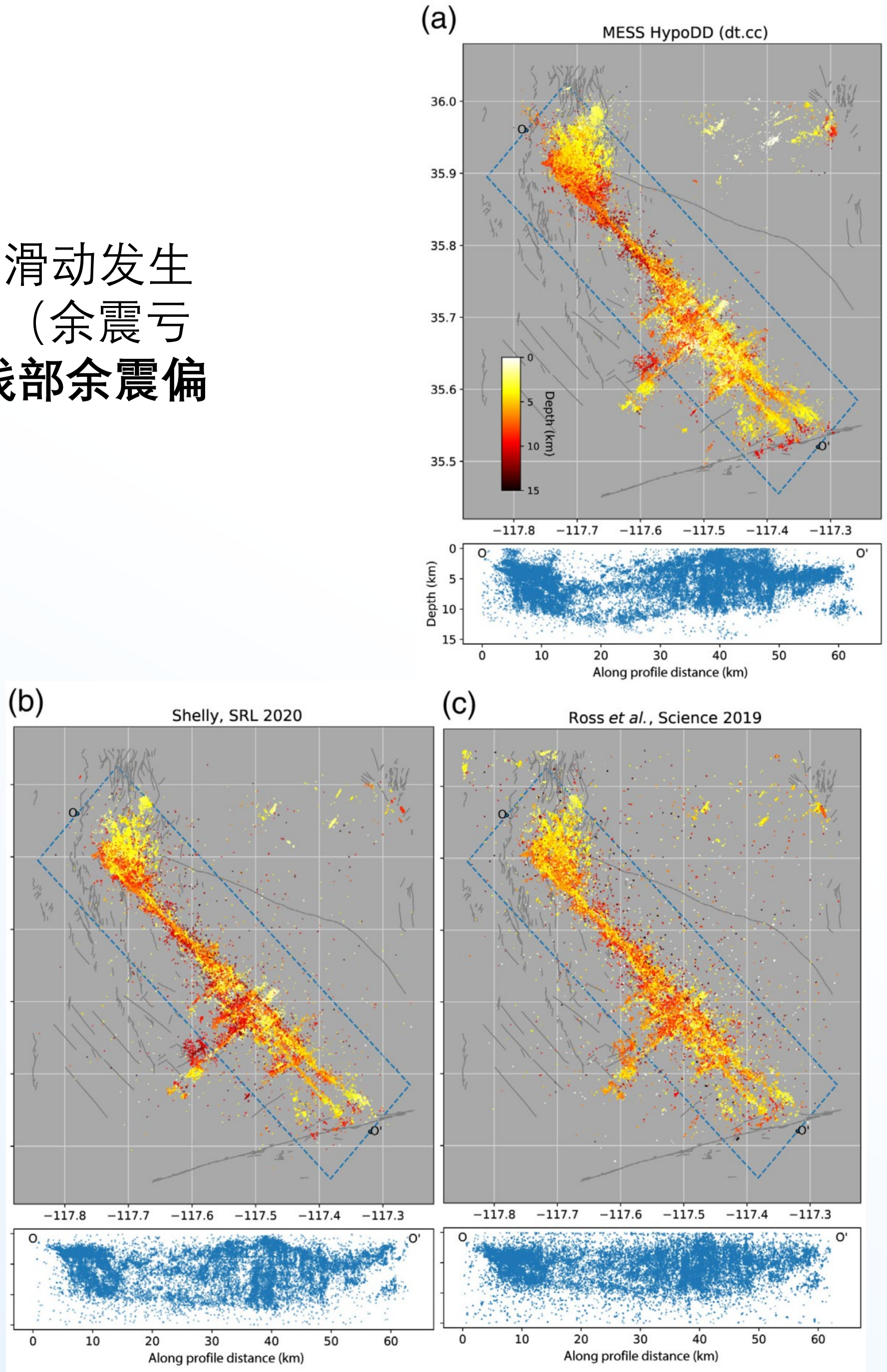
1. MESS HypoDD：该区域浅部(0–5 km)余震“相对稠密”；❌
2. Shelly (2020)：该区域浅部余震 “较为稠密”；
3. Ross et al. (2019)：在 15–25 km、0–5 km 的浅部余震明显减少，表现出“余震亏损”。❌

Claude-Opus-4.1最终答案 ❯❯❯

Ross et al. (2019)

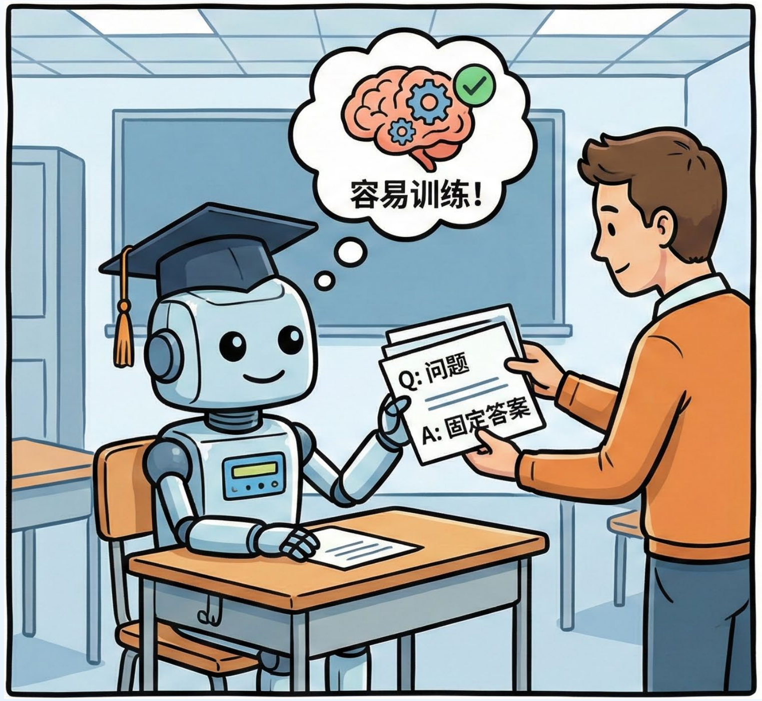
正确答案 ❯❯❯

MESS HypoDD。西北区域浅部 (0–5 km) 余震明显稀疏，甚至在 5–10 km 也偏少。

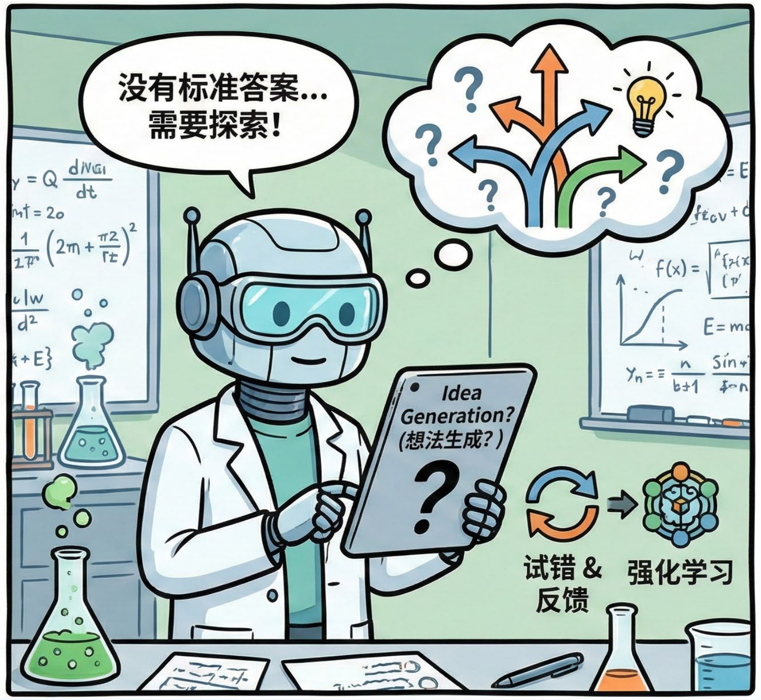


Test Time Reinforcement Learning

传统场景：有标准答案



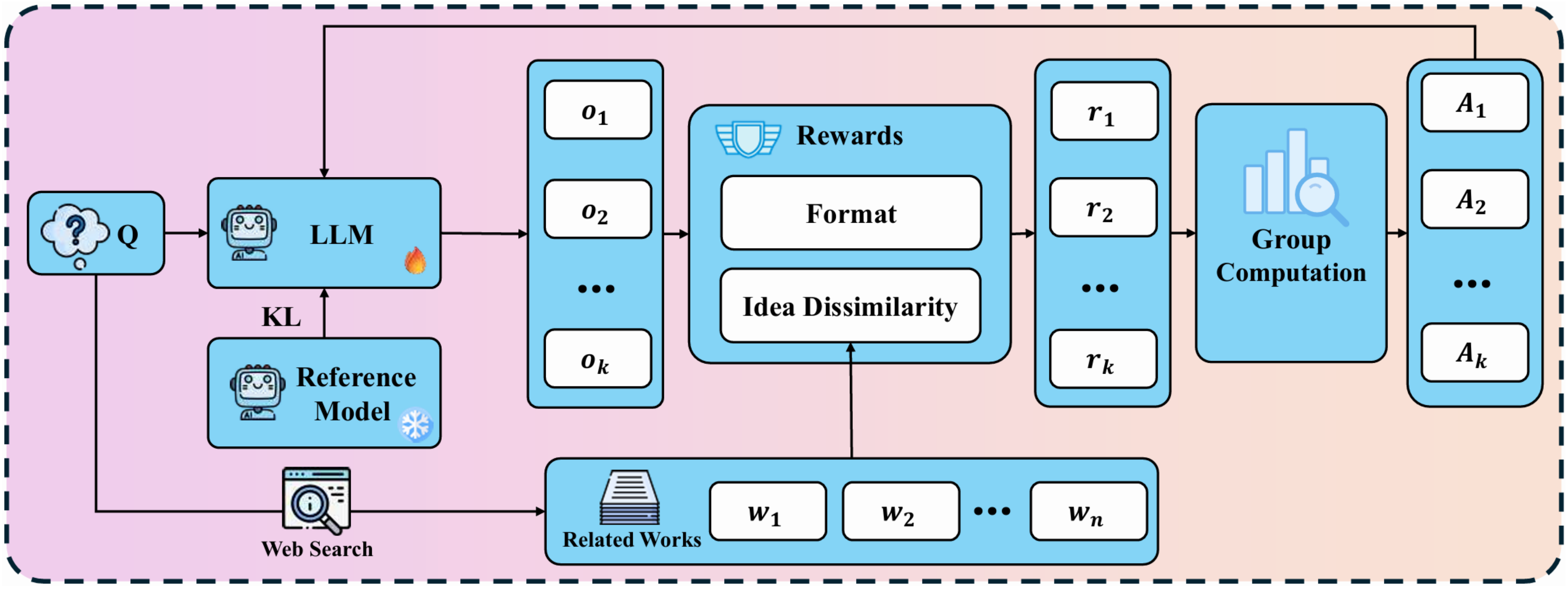
科研探索：没有标准答案



为什么要做TTRL?

1. 传统数据的答案是清晰定义的，科研训练；
2. 真实科研场景没有答案，例如idea generation。

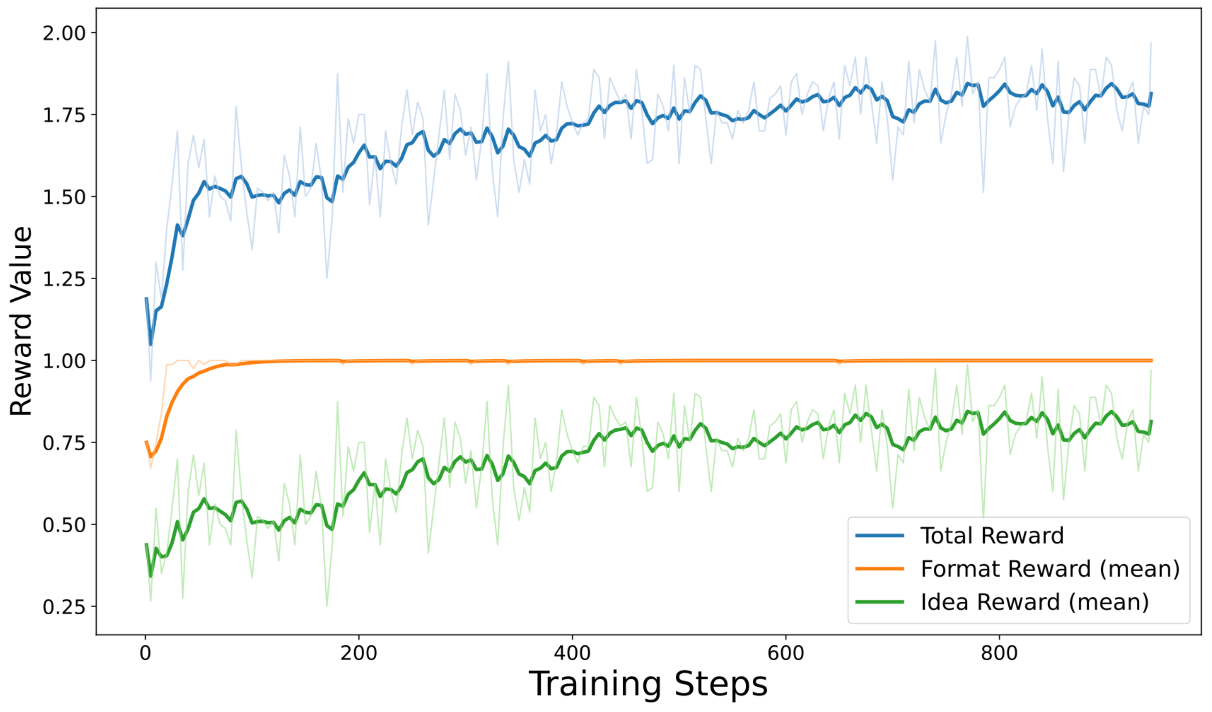
Test Time Reinforcement Learning



1. **在线检索相关文献** → 计算语义相似度 → 构造新颖性奖励（**越不相似越高分**）；
2. 使用 GRPO 优化策略;
3. 基座模型：Qwen3-8B（开源）。

Test Time Reinforcement Learning

$$R(o) = R_{\mathrm{format}}(o) + R_{\mathrm{novelty}}(o, \mathcal{W})$$



Before Post Training

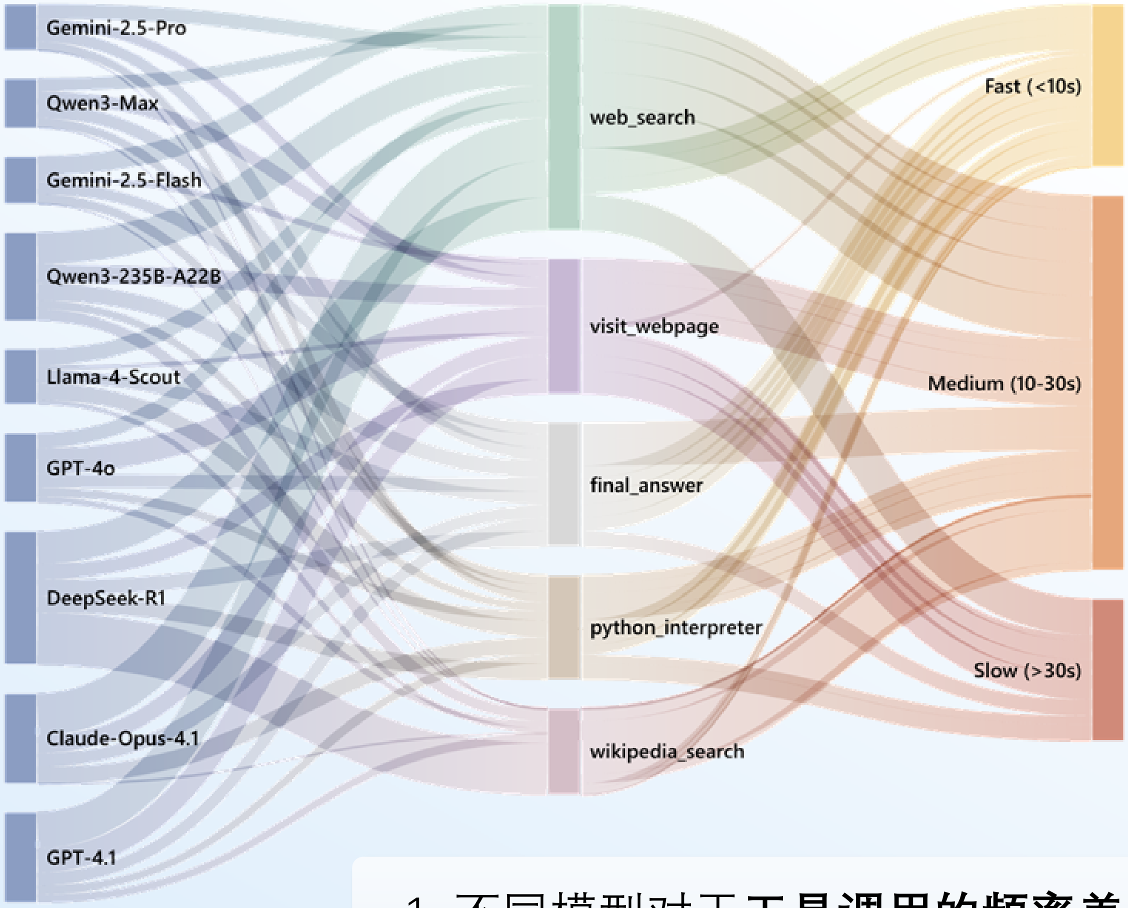
We propose a hybrid RNA 3D prediction framework integrating evolutionary signals, secondary structure priors, and physical restraints via a transformer-based architecture. This approach **combines contact map prediction with fragment assembly**, leveraging DCA-derived couplings and **Rosetta energy functions** to enhance sampling and accuracy for novel RNAs while providing reliable confidence scoring.

After Post Training

Propose a hybrid transformer-physical force field framework integrating evolutionary couplings, secondary structure priors, and physics-based scoring. The model uses a **dual-branch transformer** to decode sequence-structure relationships while a **differentiable physics engine** enforces base pairing and stacking constraints, enhanced by a confidence-aware uncertainty module for out-of-distribution detection.

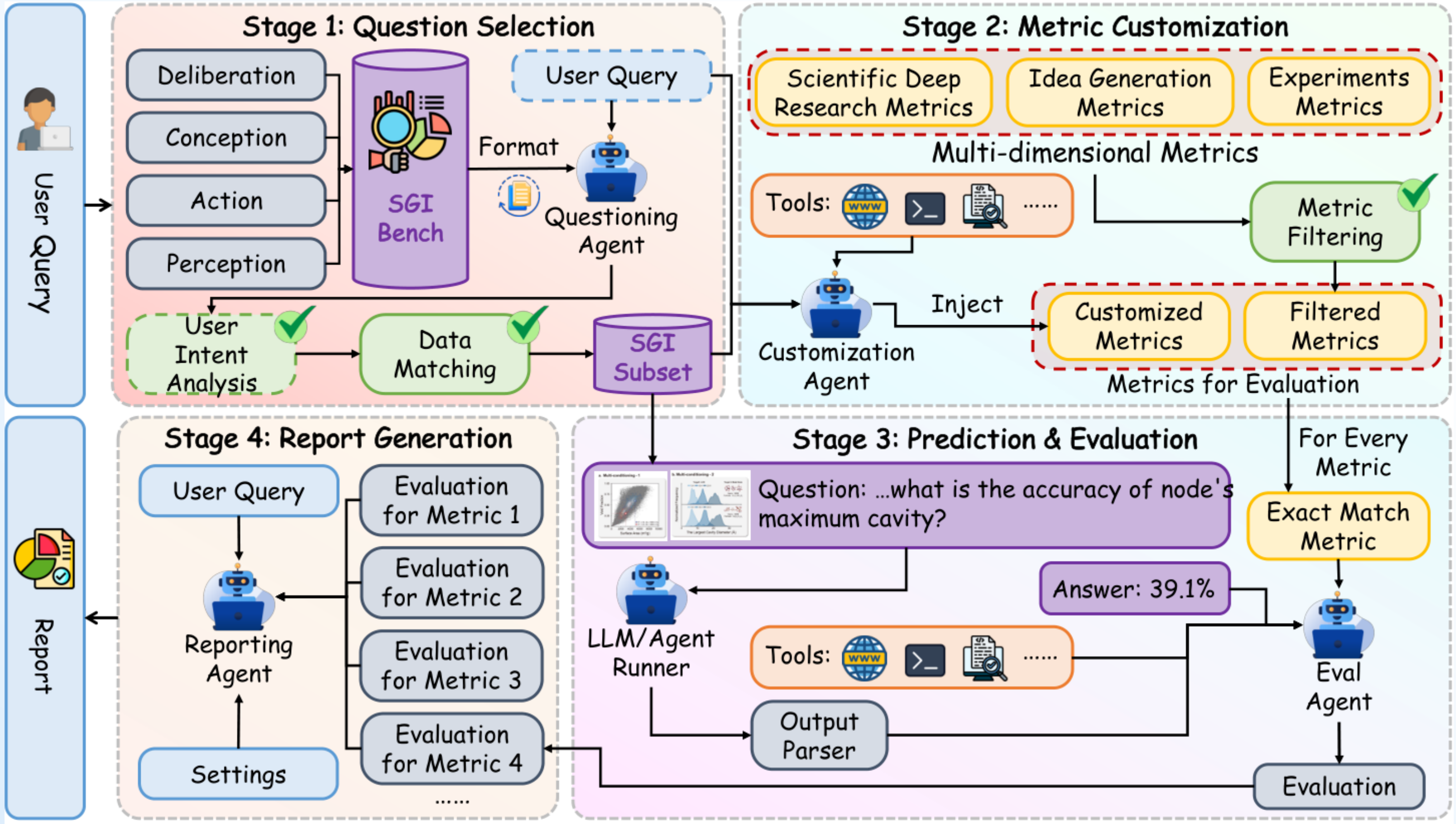
无需标准答案，新颖度由 49.36 提升至 62.06；生成结构更具体，从“拼装套路”走向“结构化创新”。

Agent Tool Integrated Reasoning



1. 不同模型对于**工具调用的频率差异显著**：其中Gemini-2.5-Pro较少调用工具，而DeekSeek-R1等模型会频繁调用工具；
2. 不同工具被使用的频率差异交大，多部分模型主要以**网页搜索作为主要的工具**；
3. **减少重复检索，提高查询质量**有助于提升回答质量的同时加速推理过程。

评测框架：Agentic Evaluation Framework



1. **Question Selection**：按学科、任务类型、样本规模与难度筛选题目；支持"覆盖全部/抽样/指定主题"等多策略；
2. **Metric Customization**：在预置科学家对齐指标基础上，结合用户意图生成"可执行评分规约"，并与内置度量合并；
3. **Predict and Eval**：模型推理与指标评测；
4. **Report Generation**：自动汇总"总体/分学科/分任务/分维度"结果，生成榜单、雷达图/柱状图。

用户自定义评测

User

I want to know which model has the best rigor in idea generation across science disciplines.

Questioning Agent

Selected idea-generation questions across Physics, Chemistry, Biology, and seven other disciplines from SGI-Bench to reflect cross-disciplinary scientific ideation.

Customization Agent

User-customized metric: Rigor (scientific strictness). Scoring criteria: 1) Logical self-consistency and tight structure (Problem→Hypothesis→Method→Metrics→Expected result); 2) Solid theory & literature grounding (properly inherits/challenges prior work); 3) Precise, verifiable problem definition (clear goals/constraints/metrics/success conditions); 4) Deep integration with research background (terms used correctly and aligned with domain definitions); 5) Methodological soundness and reproducibility (executable steps, IO loop, rationale for design choices); 6) Risk awareness & scientific criticism (bias/assumption risks/failure modes, conservative conclusions). Major deductions: missing reasoning chain, unstated key assumptions, terminology misuse, vague metrics, misaligned inheritance of background.

Evaluation Agent

Running inference and scoring...

Reporting Agent

We evaluate models on the user-defined Rigor metric for idea generation. Observations from the chart:
1. Closed-source LLMs generally achieve higher rigor; top performers cluster around the GPT-5/o3/o4-mini families.
2. Intra-family improvements are evident (e.g., Qwen3-Max > Qwen3-235B-A22B; GPT-5 > GPT-4.1), indicating sensitive capture of technical iteration.
3. Open-source progress is notable: Qwen3-Max reaches parity with GPT-4.1 on rigor.
4. Heterogeneity persists: Grok-4 is comparable to GPT-4.1 and Qwen3-Max, underscoring multi-dimensional trade-offs across instruction following, creativity, and methodological soundness.
Rigor foregrounds methodological soundness and internal coherence—models that score higher propose more structured, verifiable plans rather than purely fluent narratives.

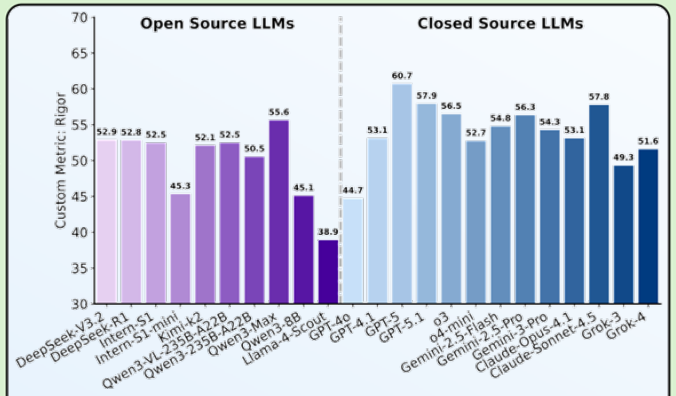


Figure: Rigor of Ideas Generated by Different LLMs

高效
每个人的科研方向不同
快速进行部分评测

定制化
每个人的评测重点不同
制定差异化指标

可读性
生成有依据，可追溯的评测
报告，方便阅读理解

未来方向

SGI

深度研究
强化证据聚合与数值鲁棒性，提升深层研究准确性

创意生成
引入规划感知与结构化监督，保障创意可行与执行细节完备

代码生成
训练需超越语法，聚焦数值分析先验与算法稳定性

湿实验协议
结合状态模拟，重点解决时序逻辑与复杂分支

多模态推理
通过细粒度视觉定位与对比训练，提升比较推理精度

测试时学习
优化多目标科学奖励体系，平衡新颖性、严谨性与安全性

欢迎关注

https://internscience.github.io/SGI-Page/

🌐 主页

📜 论文

💻 代码

🗃️ 数据

科学评测工具集
SciEvalKit

通用评测工具集
VLMEvalKit

社区交流群

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks