



Scientists AI JAMs

Franck Cappello, Argonne National Laboratory
Rio Rokota, Mohamed Wahib Institute of Science Tokyo, Riken

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

End-to-End Eval: ~~1000~~ 1500 Scientists AI JAI in 9 Labs Simultaneously (Feb.28, 2025)



Researcher participation and contributions on a voluntary basis.

1,000 Scientists Jam Session: In numbers

Researcher participation and contributions on a voluntary basis.



Total:

2800+ problems

15000+ assessed prompt responses

Argonne:

720 problems

2500 prompts



Argonne

1,000 Scientists Jam Session: Domains

Researcher participation and contributions on a voluntary basis.



Literature/Data

- Literature search, analysis, survey
- Data analysis and forecast, interpolation, extrapolation, **classification** (Point Cloud, signal, protein sequences, files, etc.)
- Anomaly detection
- Signal Analysis
- Scientific Visualization

Coding

- **Algorithm design/optimization**
- Automatic **code generation/refactoring**
- Code **translation**
- **Debugging codes** (sequential, parallel)
- Automatic code performance tuning/optimization
- **Identifying performance bottlenecks**

Experiments

- Automatic tuning of instruments
- **Experimental Design** (including autonomous workflow)
- Dark mater experiment design

Bio

- **Understanding mechanisms of Cancer**
- Understanding radiation effects on human cells
- Predictive Genomic Models

AI

- **Domain specific LLMs/Agents** (use LLMs as foundation models)
- Hyper parameter exploration for DL training.

Physics

- Battery design
- Chemical Mechanisms
- **Physics beyond standard model**

Infra.

- **Infrastructure modeling** and resilience
- Natural Disaster assessment

Math

- Surrogate model
- **Mathematical derivations**
- PDE solving
- **Convergence proving**
- Equation validity testing
- Derivative analysis
- Uncertainty estimation
- **Inverse problems**
- Statistical modeling

1,000 Scientists Jam Session: Problem Types



Researcher participation and contributions on a voluntary basis.

- Literature search, analysis, survey

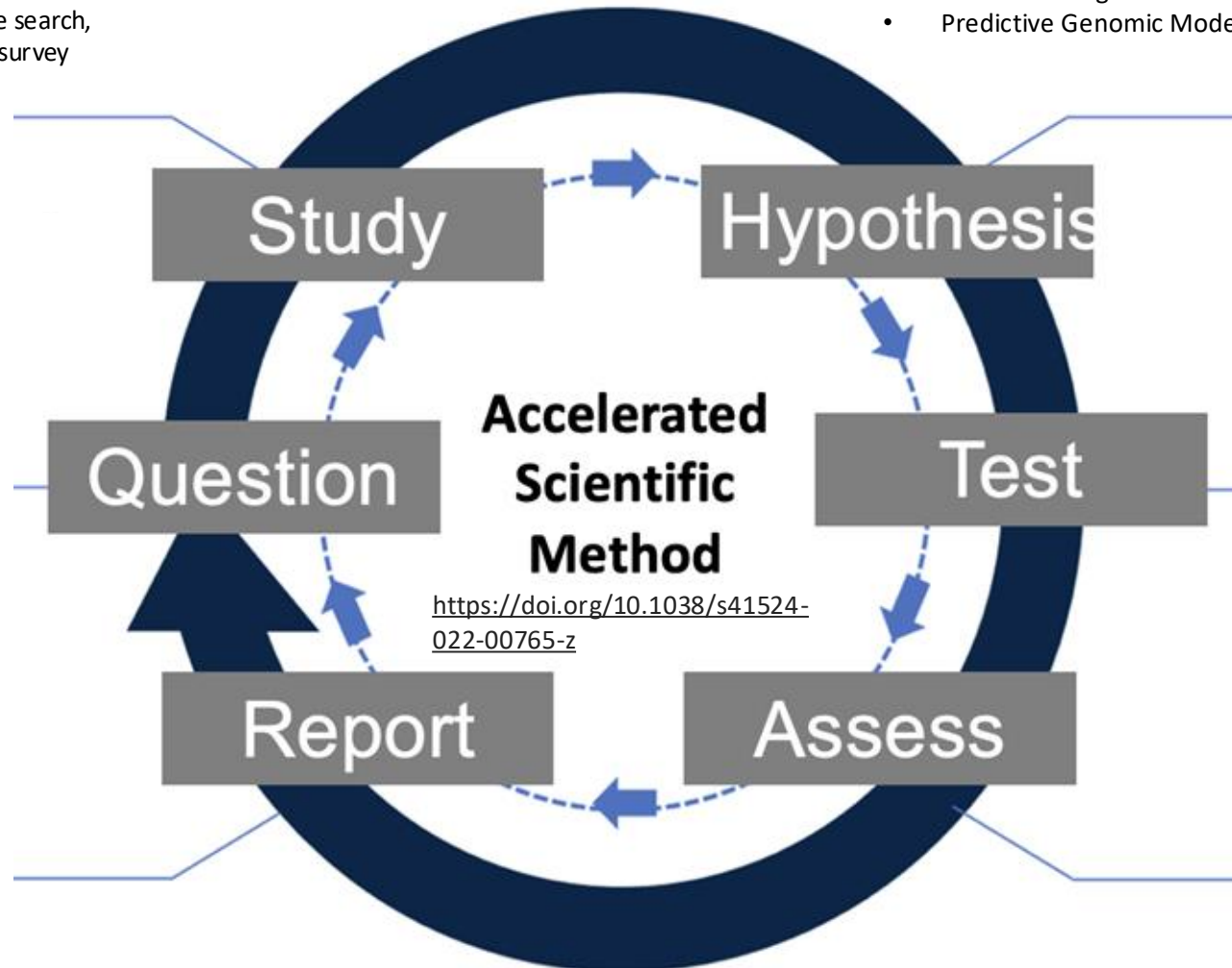
- Battery design
- Chemical Mechanisms
- Physics beyond standard model
- Understanding mechanisms of Cancer
- Understanding radiation effects on human cells
- Predictive Genomic Models

- Algorithm design/optimization
- Automatic code generation/refactoring
- Code translation
- Debugging codes (sequential, parallel)
- Automatic code performance tuning/optimization
- Identifying performance bottlenecks

- Surrogate model
- Mathematical derivations
- PDE solving
- Convergence proving
- Equation validity testing
- Derivative analysis
- Uncertainty estimation
- Inverse problems
- Statistical modeling
- Domain specific LLMs/Agents (use LLMs as foundation models)
- Hyper parameter exploration for DL training.

- Automatic tuning of instruments
- Experimental Design (including autonomous workflow)
- Dark mater experiment design

- Data analysis and forecast, interpolation, extrapolation, classification (Point Cloud, signal, protein sequences, files, etc.)
- Anomaly detection
- Signal Analysis
- Scientific Visualization
- Natural Disaster assessment
- Infrastructure modeling and resilience



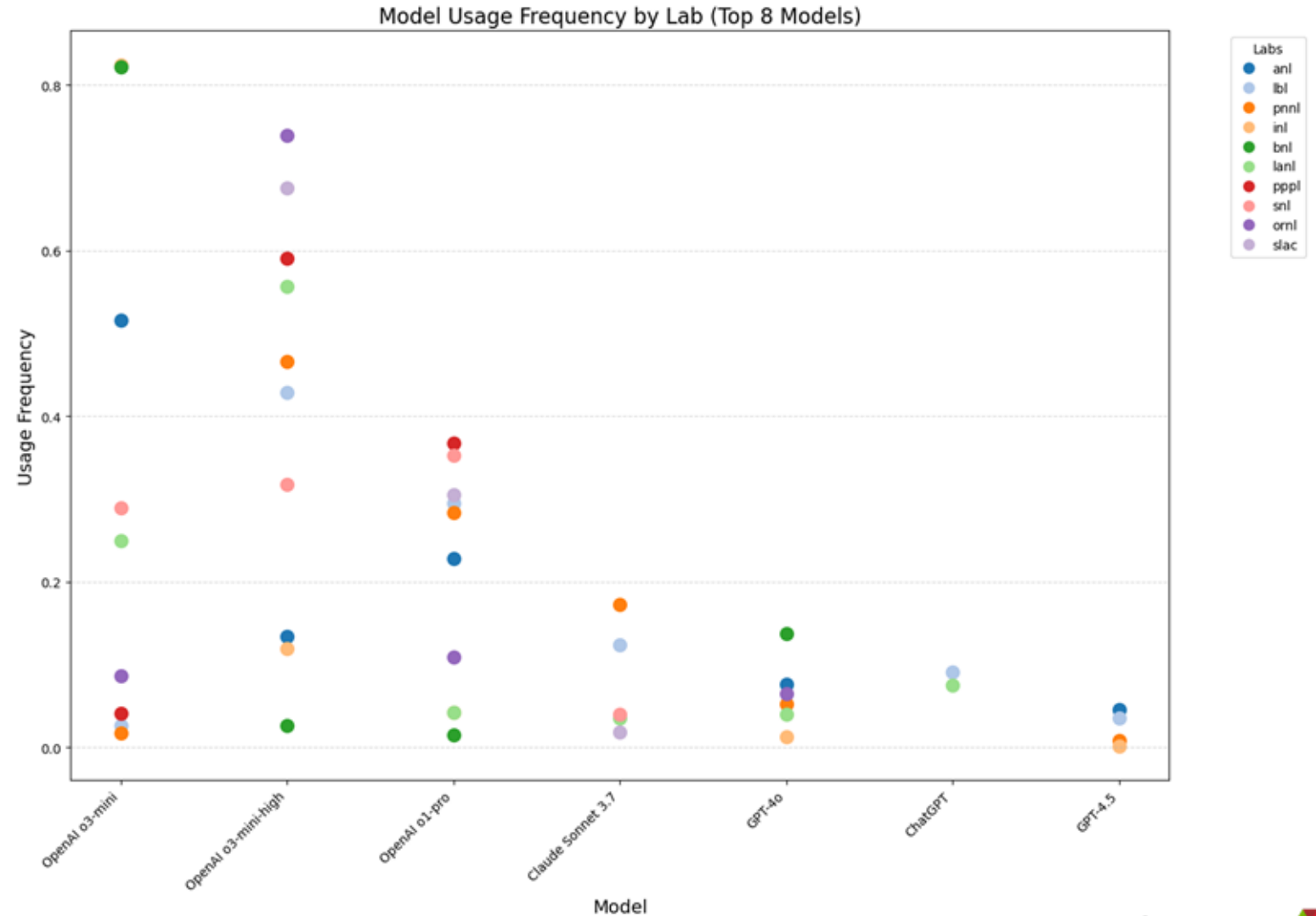
- ???

1,000 Scientists Jam Session: Model usage

Researcher participation and contributions on a voluntary basis.

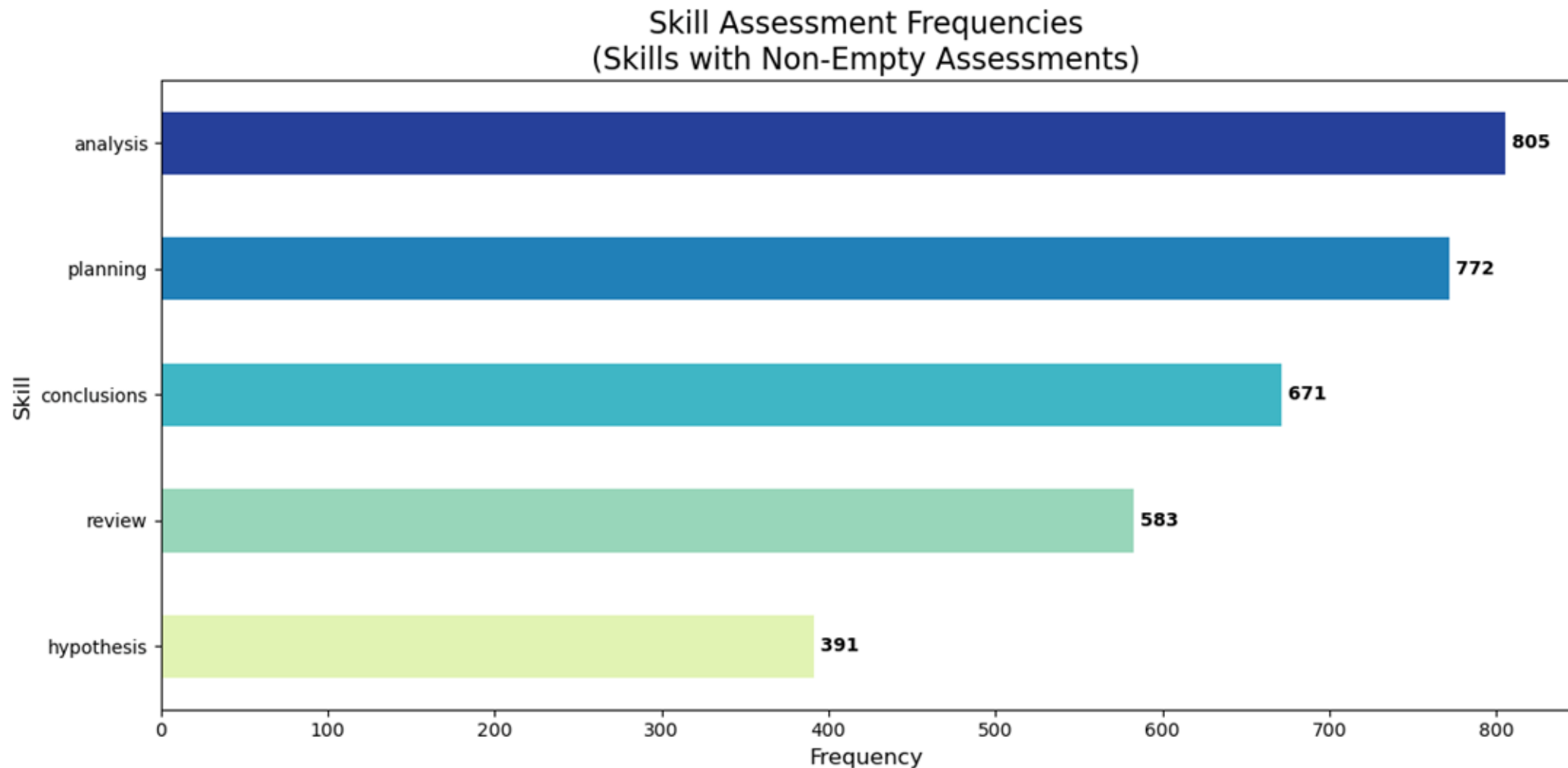


Model	Count
OpenAI o3-mini	2139
OpenAI o3-mini-high	1390
OpenAI o1-pro	1189
Unclear	380
Claude Sonnet 3.7	255
GPT-4o	253
GPT-4.5	145
ChatGPT	126
GPT-4	122
Claude Sonnet 3.7 Thinking	96
OpenAI DeepResearch	70
OpenAI o1-mini-high	51
Deep Research	47
OpenAI o1	30
Claude 3.5	19
Other	8
Free Version	6
Google Flash Thinking	2



1,000 Scientists Jam Session: Skill tested (selected by user)

Researcher participation and contributions on a voluntary basis.

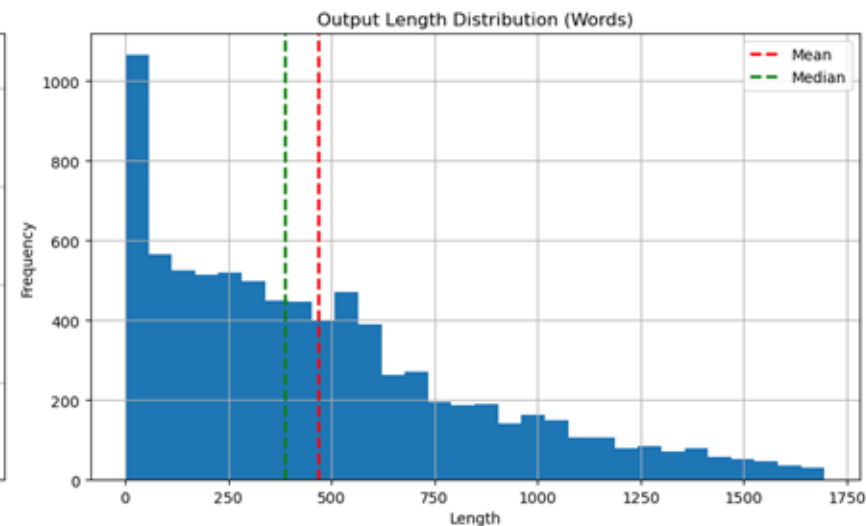
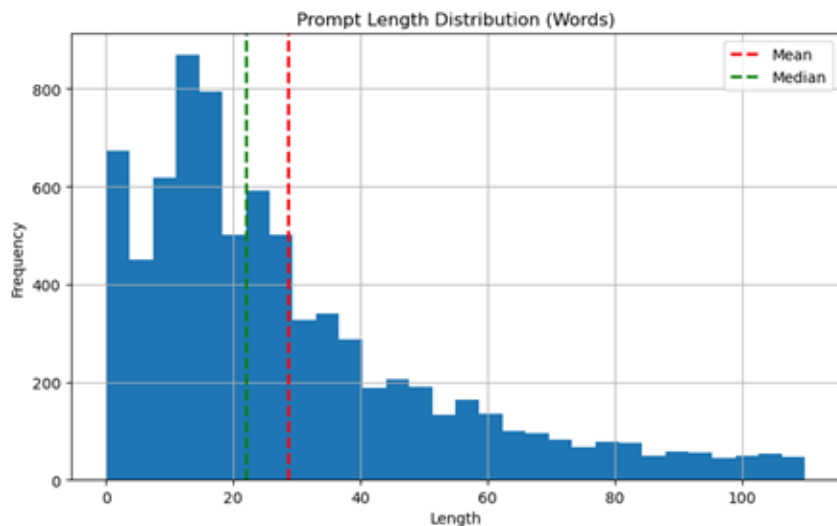


Input/Output prompt length distribution (words, tokens)

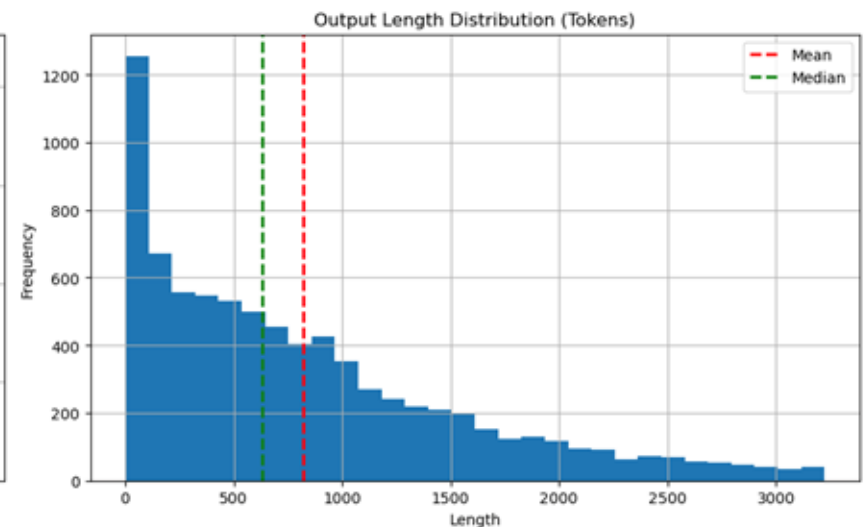
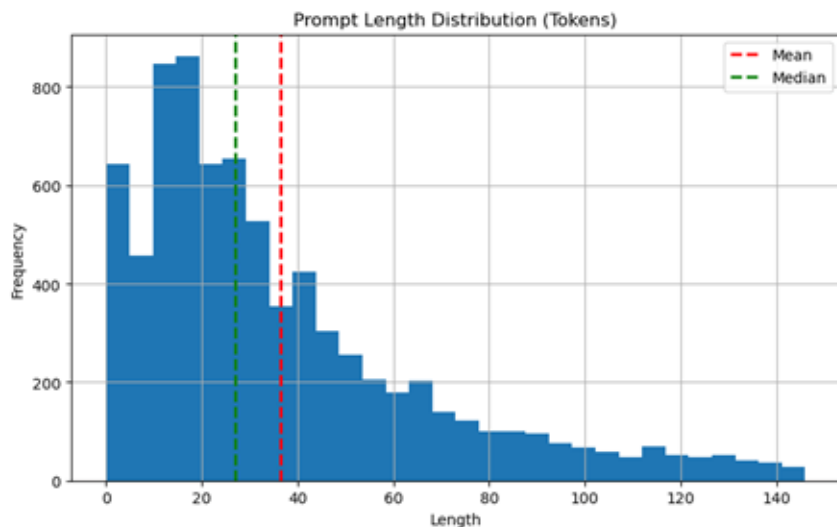
Input

Output

Words



Tokens



1,000 Scientists Jam Session: skills strength (average over the whole corpus)



LLM as a judge to automatically score (1-5) the LLMs responses

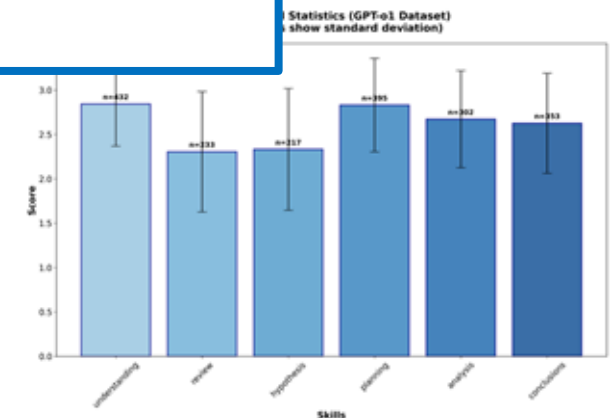
Lessons learned:

- Field style experiments are just nascent
- We need to understand better what to capture (scoring)
- We need to understand better how to diagnose AI model weaknesses based on human-AI model interactions in science context.



different skills

Result robust against change the judge model (gpt 4o -> gpt o1)



Lessons learned Communicated to Riken

Executive summary

1. **Designated Points of Contact (POCs) are Mandatory:** Each national laboratory must identify Tech Lead, Logistics Lead, Run-of-Show Lead, Review Lead, and Comms Lead POCs, ensuring clear roles and responsibilities.
2. **Consistent Communication is Critical:** Weekly sync meetings are mandatory for all POCs. Increase of meeting frequency as the date of the JAM approaches.
3. **Standardized Communication and Storage Tools:** All participants must commit to using the agreed-upon communication (Slack, email) and document storage (Box) methods.
4. **AI Lab Model Providers Have Specific Commitments:** These include providing a POC, conducting a training webinar, and ensuring onsite technical support during the Jam.
5. **Streamlined Model Access is Recommended:** Utilizing IP address-based access for large AI Jams is suggested to simplify account management.
6. **Hosting Labs Hold Significant Logistical Responsibility:** They are responsible for all onsite logistics, participant communication, registration, and disseminating crucial guidance.
7. **Legal and Compliance Aspects Must Be Addressed:** Timely communication and documentation of IRB requests (Human Subject Research) and HQ-GC exemption requests are essential.
8. **A Common Data Collection Tool is Required:** All labs must use the same tool for data collection. Argonne has used a collection tool during the JAM and identified improvements planned for future iterations.
9. **Thorough Corpus Review is Necessary:** Each lab needs a designated POC for corpus review, involving technical expertise and input from legal, security, and other relevant teams, with a two-week review commitment.
10. **Quality Control and Timely Distribution of Reviewed Data are Key:** The collecting lab is responsible for corpus quality, and the aggregated, validated corpus must be promptly distributed to participants.

Collection Tools Shared with Riken

Experience

Problem Description

Problem Setup

On average, what is your level of experience with advanced AI systems such as ChatGPT 4o, Claude, LLama3, etc? If answering for a team, provide the level of experience of the most experienced team member.

- I use them several times a day
- I use them several times in a week
- I have never used them before or used them infrequently

On average, what is your level of experience with advanced reasoning AI systems such as O1, O3, Gemini 2.0, Perplexity-Pro-Reasoning? If answering for a team, provide the level of experience of the most experienced team member.

- I use them several times a day
- I use them several times in a week
- I have never used them before or used them infrequently

What model did you use?

ChatGPT o3 Mini

Title for your experiment *

What is your overall research goal/objective for this experiment?

Describe the problem in a paragraph or more *

How would you describe the level of difficulty of the problem?

How realistic (true to life) is the problem you will work on today?

Please provide any additional information you consider relevant

I certify that this problem does not contain any controlled unclassified information, information subject to export controls or personally identifiable information (PII)

Start Prompting

Prompting

First, think about your prompt

What is the goal of this prompt?

What is the prompt?

Please upload any files files provided in the prompt if any (limit 50MB/file)

Click to upload files

Or URL for larger datasets containing files larger than 50MB. Do not provide both FILES and FILES URL

Run the prompt in AI Interface Website in another tab or window

Please copy and paste the output

What skill(s) did you explore with this prompt?

Introduction to Skills Evaluation Problem Understanding Literature Review Hypothesis Generation Planning/Design
Result Analysis Generate Conclusions Other Tasks

Please explore any skills you think are appropriate for this response by clicking on the tabs for the skills you wish to explore. You may scores as many or as few as you like for each response. When you are finished, you can click "keep prompting" ask a new prompt or "finish" to move onto a new evaluation

I certify to the best of my knowledge that this problem does not contain any controlled unclassified information, information subject to export controls or personally identifiable information (PII)

Researcher participation and contributions on a voluntary basis.

Prompts

Attachment

Model Response

Response assessment

Riken AI JAM

Japan Scientist AI Jam Session

Online Tutorial: Nov. 11
225 participants

Dry Run: Nov. 28
50 participants

AI Jam (Tokyo): Dec. 16
124 participants

AI Jam (Kobe): Dec. 18
95 participants

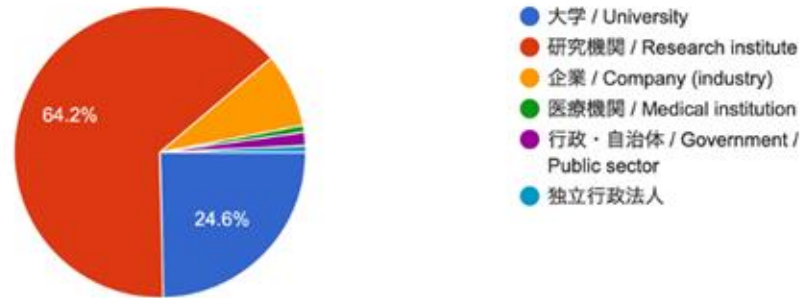
Model providers: OpenAI, Anthropic, Google, AWS, NVIDIA
Q&A through Discord
Collection tool using Chrome browser extension
RIKEN open model platform + MCP server



Post-Jam Survey

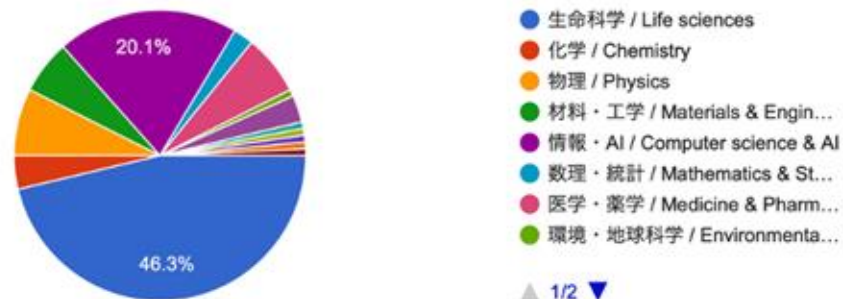
所属機関の種別（単一選択） / Type of organization (Single choice)

134 responses

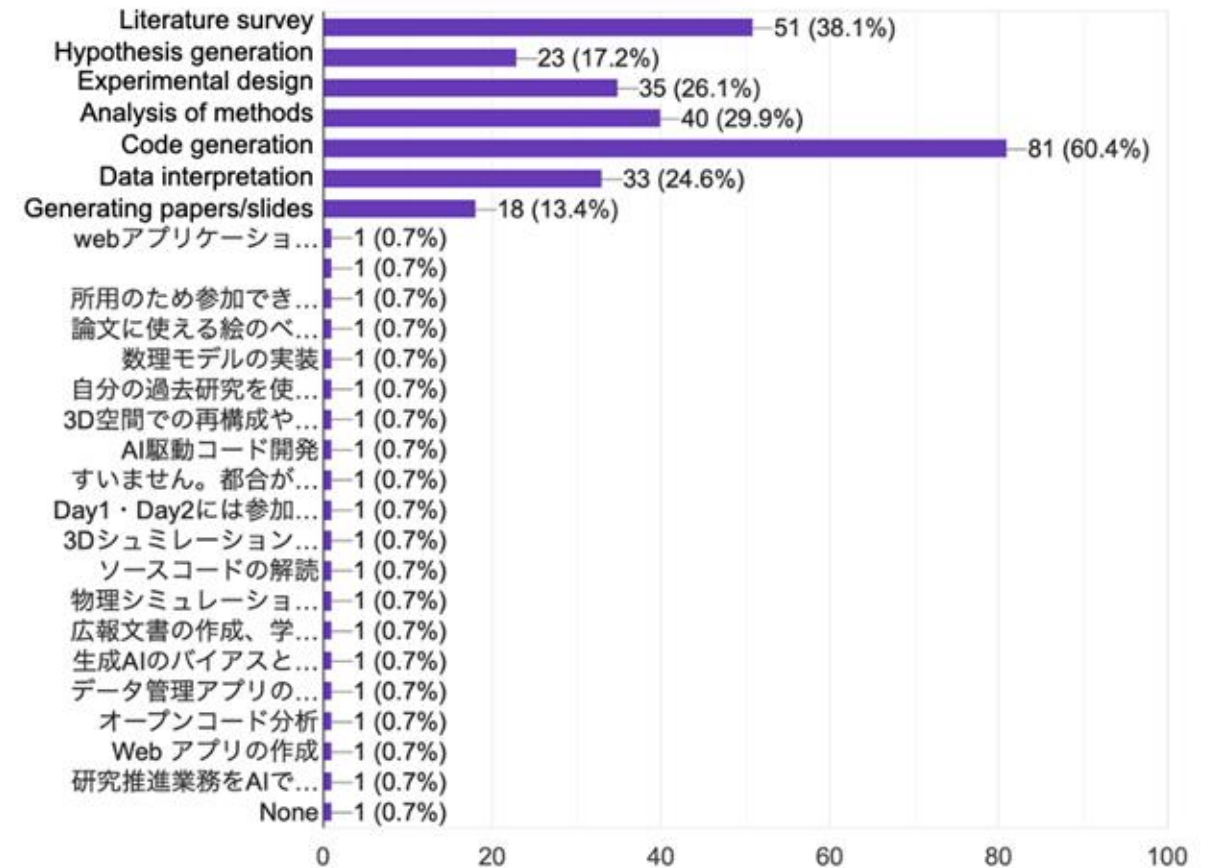


研究分野（単一選択） / Research field (Single choice)

134 responses



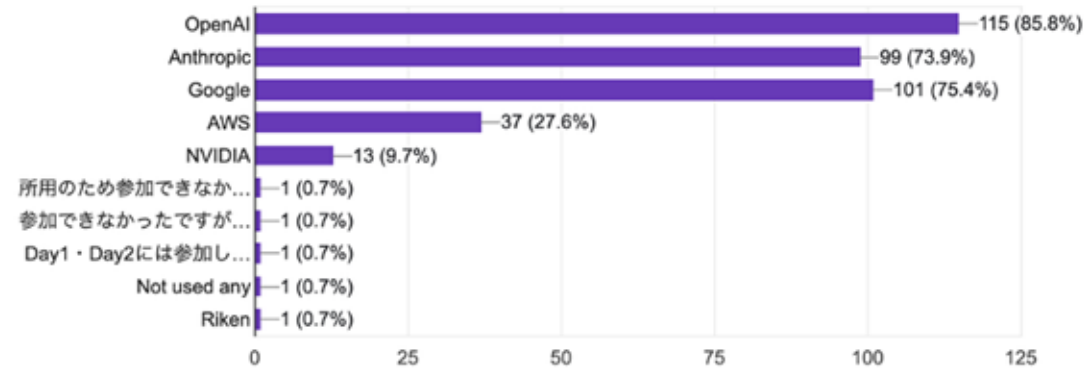
今回取り組んだ課題のタイプ（複数選択） / Task type(s) worked on (Multiple choice)



Post-Jam Survey

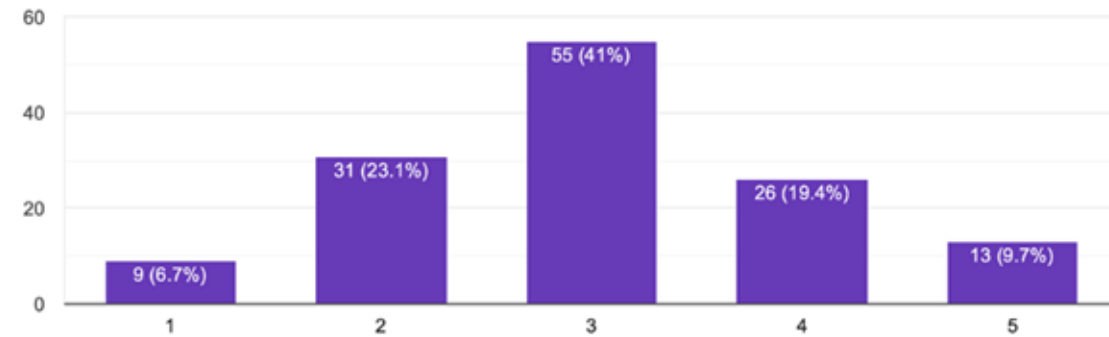
主に使用したAIモデル/プロバイダー（複数選択） / AI model/provider used (Multiple choice)

134 responses



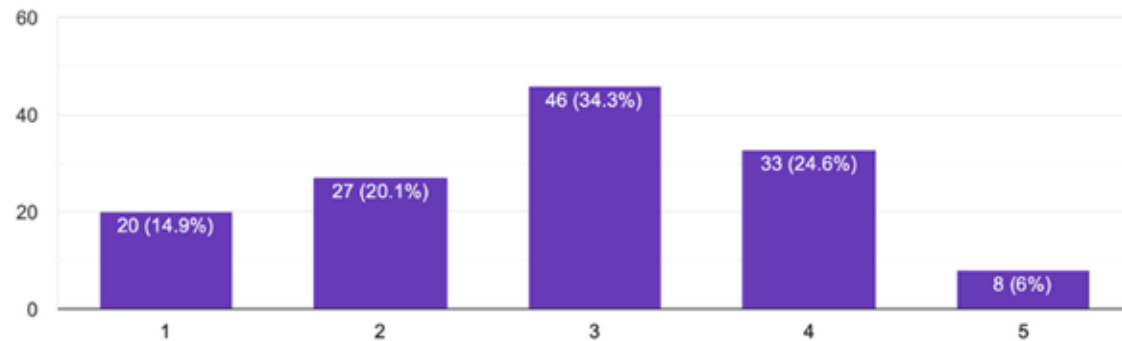
事前案内の分かりやすさ（5段階） / Clarity of pre-event communication (5-point)

134 responses



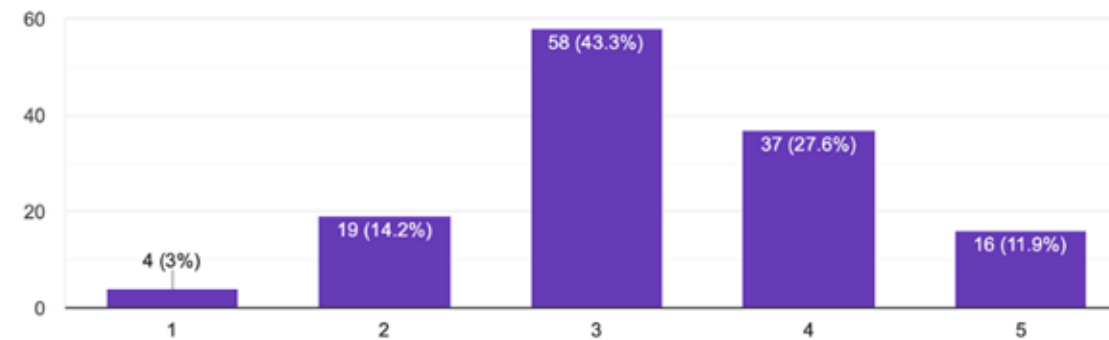
イベント前の「課題の準備状況」（5段階） / How well-prepared your problem statement was (5-point)

134 responses



オンラインチュートリアルの満足度（5段階） / Satisfaction with the online tutorial (5-point)

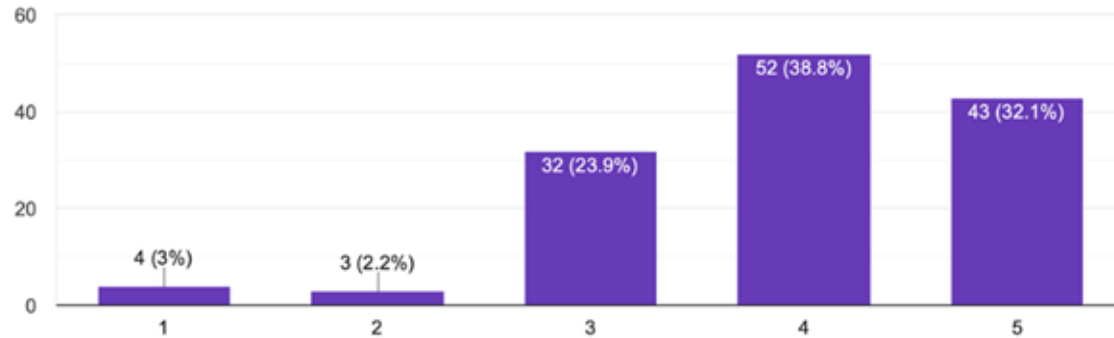
134 responses



Post-Jam Survey

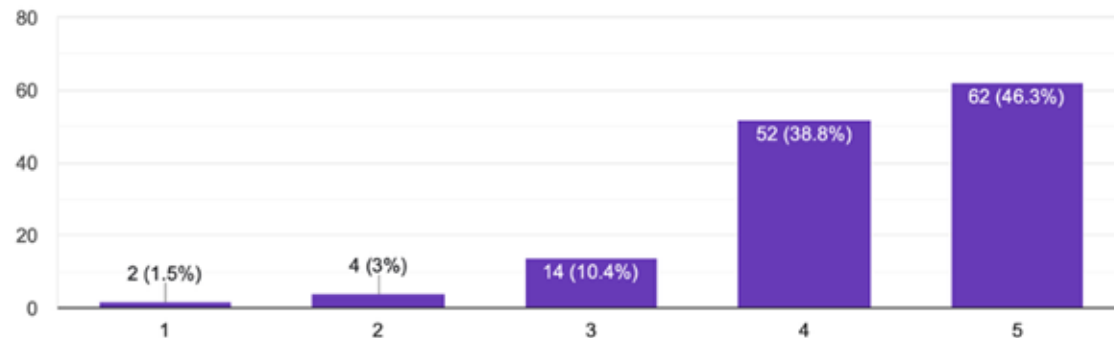
本番セッションの運営面の満足度（5段階） / Satisfaction with on-site operations (5-point)

134 responses



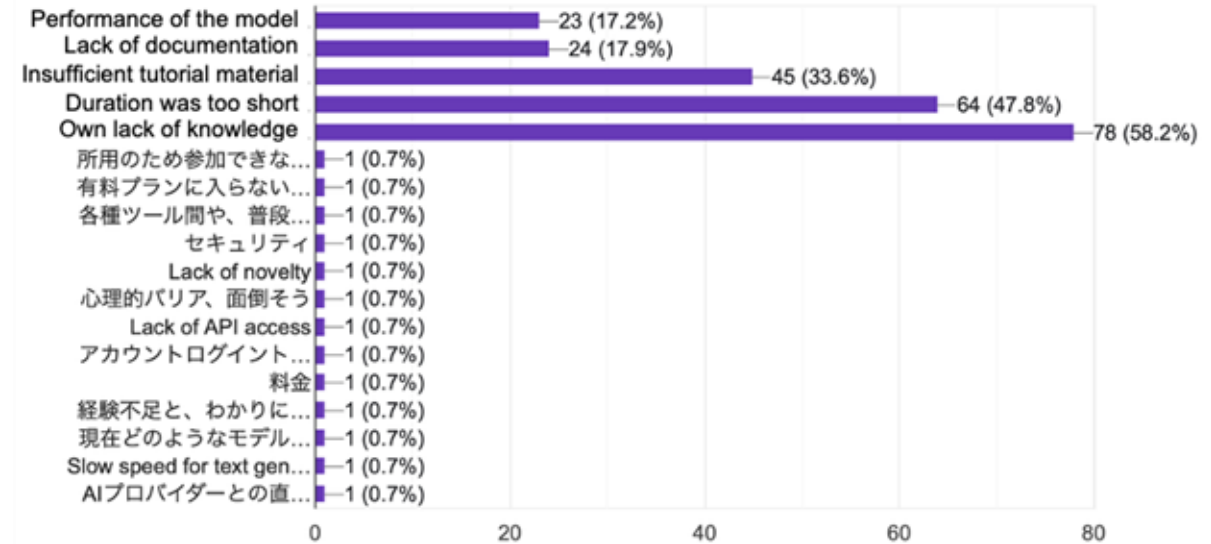
今回の体験が次の研究活動にどの程度影響を与えますか？ / Will today's AI Jam have an impact on your research?

134 responses



AIモデルやツール利用における最も大きな阻害要因は何だったか？（複数選択） / What was the biggest problem when using the AI tools? (Multiple choice)

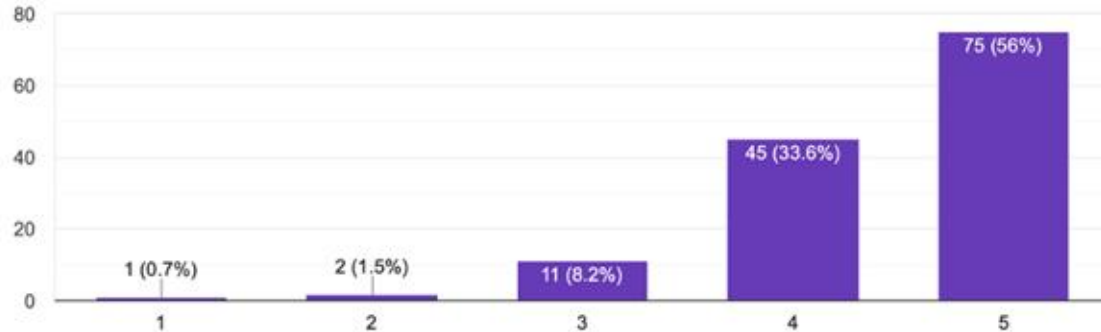
134 responses



Post-Jam Survey

またAI Jamがあれば参加したいですか？ / If there is another AI Jam, will you participate?

134 responses

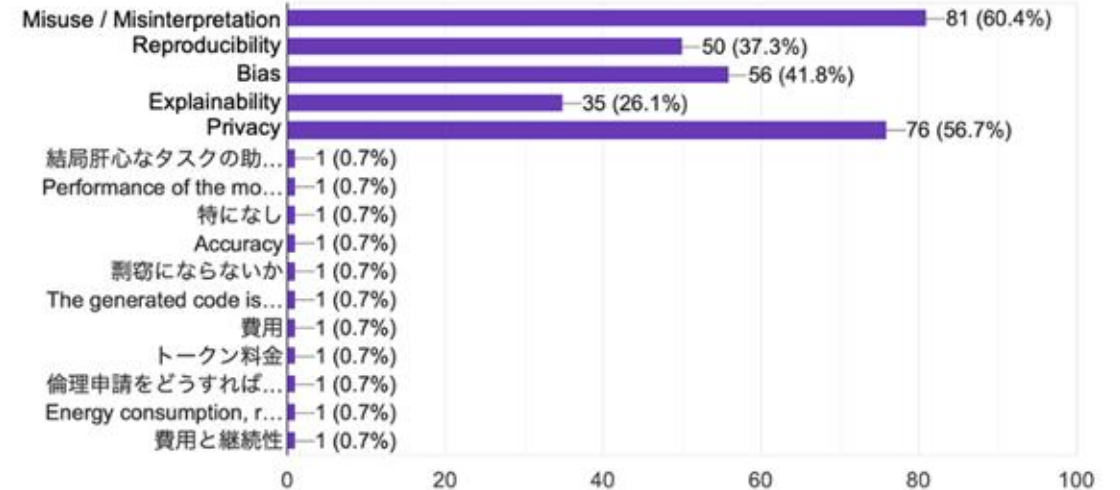


Positive feedback

- On-site support from Frontier AI companies was very helpful
- Being able to compare the latest high-end models was valuable
- Being able to use the most expensive model for free allowed me to do much more
- I learned many new features and tools that I didn't know existed
- I learned a lot through conversations with other participants
- The capability of the latest high-end models exceeded my expectation

研究でAIを使う際、最も懸念している点はどれですか？（複数選択） / What is your largest concern when using AI for research? (Multiple choice)

134 responses



Negative feedback

- Duration of the AI Jam was too short to do anything significant
- It would have been better to have more guidance than to have a free format
- The goal of the event was not clearly announced in advance
- Tutorials were too brief to actually try what was shown
- I wish more examples were shown during the tutorial
- Account setup should have been announced better/earlier

Use NotebookLM to generate pretty slides

Accelerate Scientific Discovery: An AI-Powered Workflow

MEET YOUR AI RESEARCH TOOLKIT

Gemini Pro: The Creative Partner

An "open-world" AI for brainstorming, coding, writing and general-purpose questions.



NotebookLM: The Focused Expert

A "closed-world" AI that becomes an expert only on the research papers you provide.



Agents (Deep Research & Idea Generation): The Specialist Team

Autonomous agents that perform complex tasks like exhaustive literature searches or novel hypothesis generation.



Code Assist & CLI: The Developer Tools

Integrated tools for writing code, managing files, and automating tasks in your IDE and terminal.



5-PHASE AI-POWERED RESEARCH WORKFLOW

PHASE 1: BROAD EXPLORATION & LITERATURE REVIEW

Use **Gemini** for initial brainstorming and **Deep Research** for comprehensive literature synthesis.



PHASE 3: EXPERIMENTATION & DEVELOPMENT

Build simulations with **Code Assist** in your IDE; manage files and scripts with the **Gemini CLI**.



PHASE 5: DISSEMINATION & WRITING

Use **Gemini** in a collaborative canvas to draft, retine, and structure your manuscript.



PHASE 2: FOCUSED ANALYSIS & HYPOTHESIS GENERATION

Upload key papers to **NotebookLM** for deep analysis and use **Mea Generation** to propose novel solutions.



PHASE 4: ANALYSIS & INTERPRETATION

Generate data visualization code with **Code Assist**; connect results back to sources with **NotebookLM**.



Carbon Natotube 3D Generator

Carbon Nanotube 3D Generator

SWCNT 3D Generator

n: 6, m: 5, 長さ(A): 40

Download

PROPERTIES

カイラリティ (n, m)	(6, 5)
タイプ	カイラル型 (Chiral)
予測カラー	#6a329e
直径 (d)	0.747 nm
カイラル角 (θ)	27.00°
円周長 (L)	2.35 nm

VIEW SETTINGS

原子サイズ

Color Info

ご提供いただいた論文「Colors of Single-Wall Carbon Nanotubes」に基づき、光吸収スペクトルをシミュレートして色を決定しています。

*単体型CNT (管径 $d < 1.5$ nm) は、可視光域の吸収ピークにより鮮やかな色 (緑色) を示します。例えば、(6,5)は紫色、(7,5)は黄緑色になります。

思考プロセスを表示

ユーザーのリクエストに応じて、生成されたカーボンナノチューブのデータをダウンロードする機能を追加します。具体的には、以下の修正と機能追加を行います。

- データ構造の管理方法の変更:
 - これまで描画 (useEffect) の中で行っていた座標計算を useMemo に切り出し、エクスポート機能からも原子座標データ (finalAtoms) にアクセスできるようにリファクタリングします。
- エクスポート機能の実装:
 - CIF形式 (.cif): 結晶構造解析ソフト (VESTAなど) で読み込める形式。仮想的な単位格子 (P1空間群) を設定し、分率座標を出力します。
 - XYZ形式 (.xyz): 多くの化学ソフトで扱える汎用的な座標フォーマット。
- UIの更新:
 - 画面右上のコントロールエリアに「ダウンロード」ボタンを追加します。

以下のコードでファイルを更新します。

一緒に文章を書いたり、コンテンツを作成したりしましょう

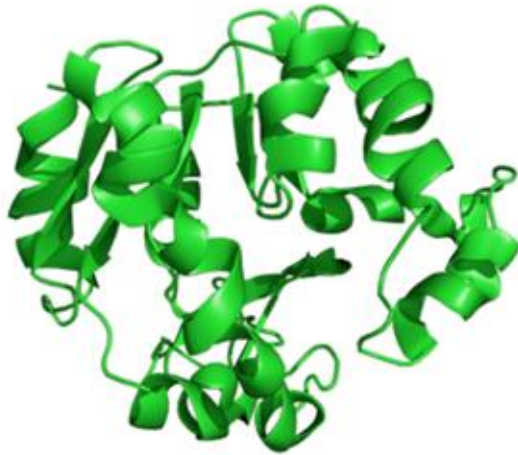
Canvas x 思考モード

stepakoff-live.com のチャットはモデルのトレーニングには使用されません。Gemini は不正な情報を表示することがあるため、生成された回答を再確認するようにしてください。アライバード Gemini

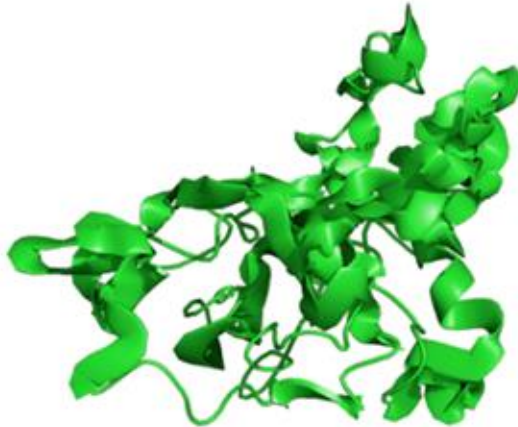
Modifying Existing Libraries for BioEmu

既存のライブラリに対する修正を加える仕事を行いました。ライブラリは参照しなければいけない要素が多いため、自分で指示するのが難しかったのですが、「AIに対する指示を作成するためにAIを使う」という方針に変えたところ、結果がよくなりました。結果として、アイデアスクラッチの状態から、適切な比較実験の設計、機械学習モデルの実装、トレーニングまでを行うことができました。

BioEmu × MSA



BioEmu × ESM



BioEmu × Boltz2



Using Claude Skills for Laboratory Automation

```
• Bash(mamba run -n forOpentrons opentrons_simulate
  "C:\Users\EliottJacopin\Documents\Career\Opentrons\pcr_setup_protocol.py")
  L === Starting PCR Setup Protocol ===
  Setting up 8 PCR reactions of 50 µL each
  ... +187 lines (ctrl+o to expand)

• Perfect! The protocol simulates successfully! Here's a complete guide for the Opentrons command-line tools:

Summary: Installation & Usage

1. Environment Setup ✓ Complete

# Create environment (already done)
mamba create -n forOpentrons python=3.11 -y

# Install Opentrons (already done)
mamba run -n forOpentrons pip install opentrons

2. Available Commands

opentrons_simulate - Simulate protocols without hardware
- Purpose: Test protocols, verify logic, estimate run time
- Doesn't require: Physical robot connection
- Usage:
mamba run -n forOpentrons opentrons_simulate path/to/protocol.py

opentrons_execute - Run protocols on real hardware
- Purpose: Execute protocols on connected Opentrons robot
- Requires: Physical robot connection (USB or network)
- Usage:
mamba run -n forOpentrons opentrons_execute path/to/protocol.py

3. Common Use Cases

Test your protocol (simulation):
mamba run -n forOpentrons opentrons_simulate pcr_setup_protocol.py

Get detailed output (with log level):
mamba run -n forOpentrons opentrons_simulate pcr_setup_protocol.py --log-level DEBUG

Check for warnings/errors only:
mamba run -n forOpentrons opentrons_simulate pcr_setup_protocol.py --log-level WARNING

Run on real robot (when connected):
mamba run -n forOpentrons opentrons_execute pcr_setup_protocol.py
```



AI can be applied to Laboratory Automation
However, the API must be publicly available.

With Claude Skill, specific functionalities and knowledge can be easily added to the model, enabling rapid prototyping for LA.
However, unlike MCP, it is not possible to tell which response triggered a Skill call.

Sharing skills within RIKEN / among researchers will help accelerate the overall progress of science.

Toward Co-scientist JAMs

Hypothesis Generation: The Next Frontier

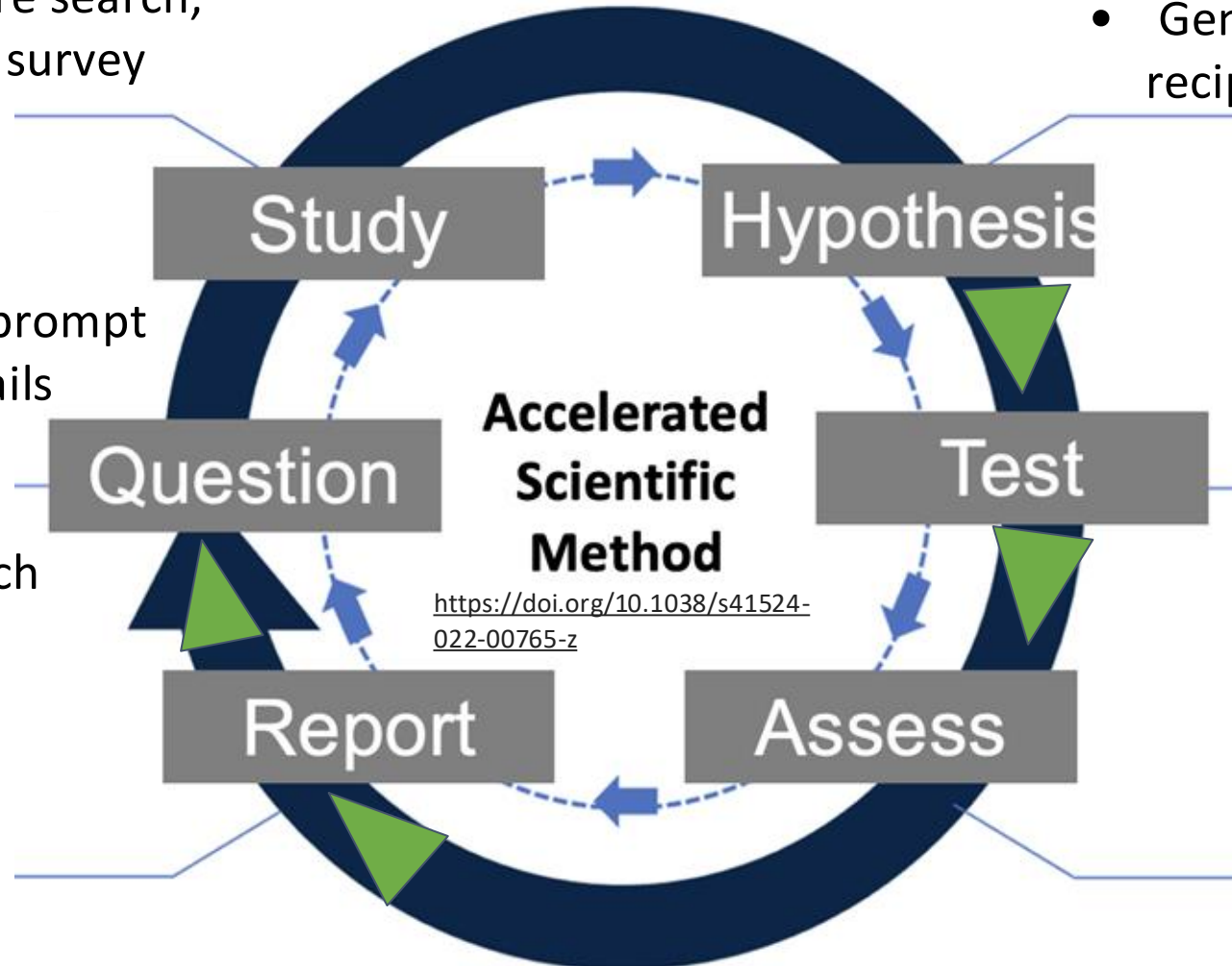
HG is at the beginning of the research workflow
It conditions everything else.

- Literature search, analysis, survey

- Generate, compare, rank ideas. Meta analysis of the generated ideas
- Generate experiment recipes

- Scientist craft a prompt with enough details and directions to avoid confusion and guide research

- ???



2025 → Co-Scientists Emergence

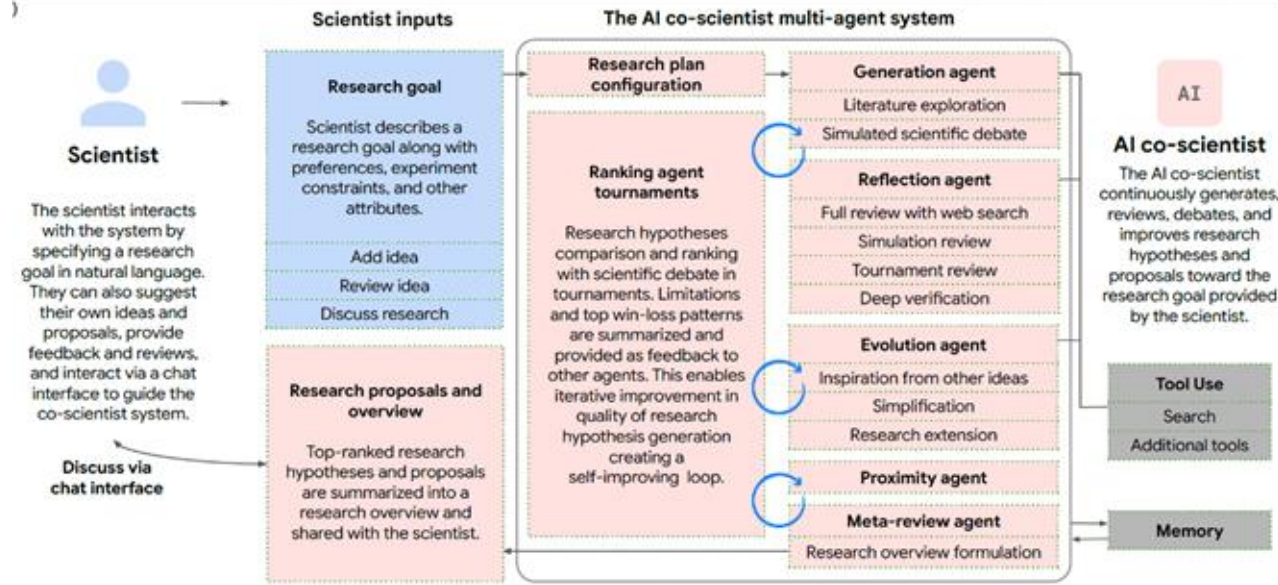
[USA] Google co-scientist Toward and AI co-scientist, arXiv:2502.18864

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)
- **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



Others:

- [USA] Rick's Scientia and Wisteria (not tested yet)
- [Japan] Sakana AI Scientist-v2: [ICLR2025 Workshop Experiment] <https://github.com/SakanaAI/AI-Scientist-v2>
- [China] Novix: <https://novix.science/> Neups 2025 Spotlight paper <https://github.com/HKUDS/AI-Researcher>
- [China] NovelSeek InternAgent: <https://github.com/InternScience/InternAgent>

Hypothesis Generation: The Next Frontier

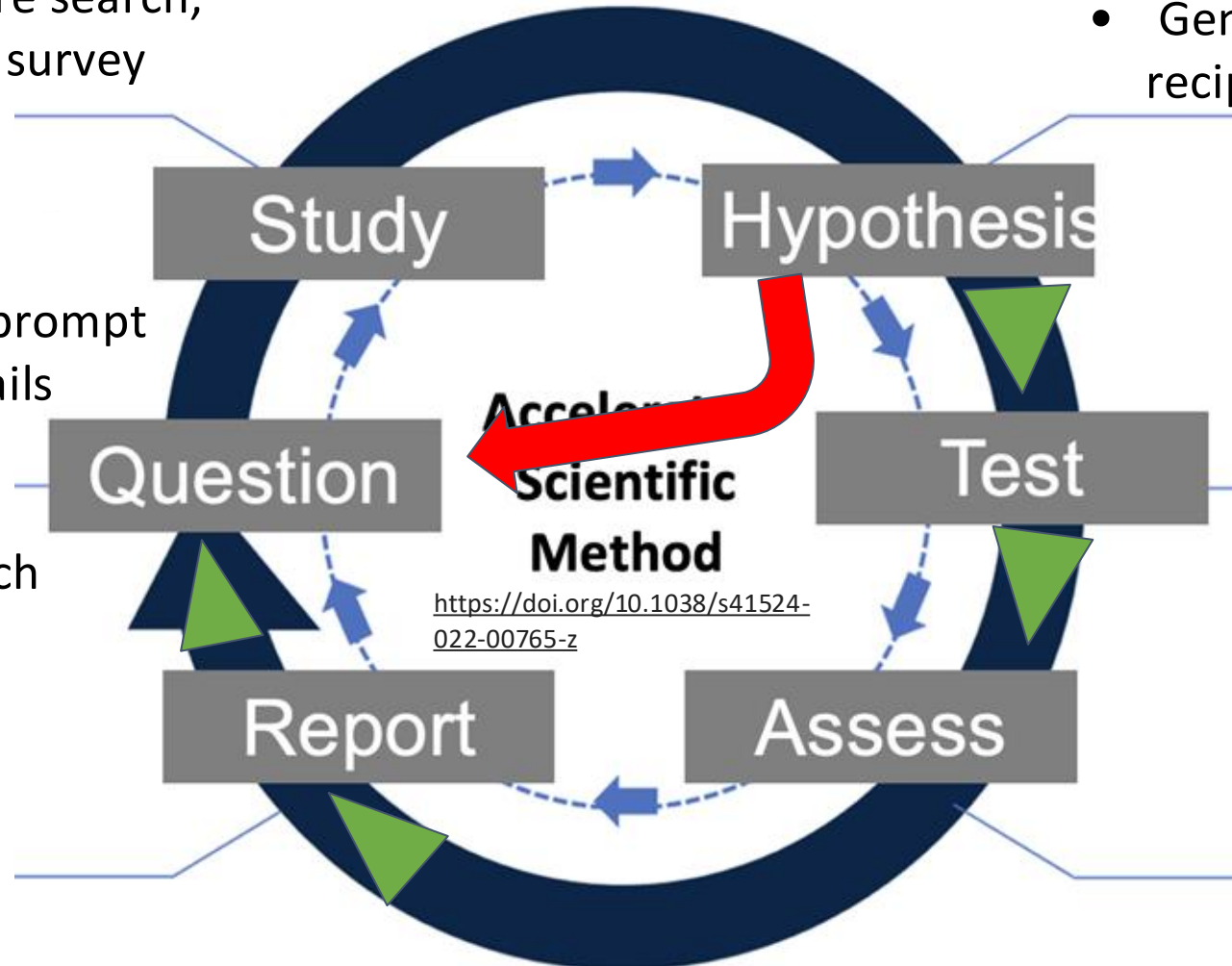
This is not the classic LLM: very sensitive to prompt structure and level of details.

- Literature search, analysis, survey

- Generate, compare, rank ideas. Meta analysis of the generated ideas
- Generate experiment recipes

- Scientist craft a prompt with enough details and directions to avoid confusion and guide research

- ???



Early feedback on AI-Researchers

- Tested multiple co-scientist systems.
- Generation can takes hours/days
- The idea generation is extremely sensitive to the prompt. It is critical for the user to state the problem to solve with as many details, constraints, and priorities as possible. The user should express all implicit assumptions concerning the scientific background, specific context of the research, preferred directions, what to prioritize in case of trade-offs, etc.
- It is important to check the tool understanding and consideration of the submitted problem.
- Once launched, it is important to be able to correct the generation (because the generation can take hours/days)
- Analysis of proposed hypothesis compared to the prompt by an LLM (Claude) recommends strengthening the hypothesis: generation is missing key details and validation steps, and priorities may not be well ranked.
- Generates sophisticated ideas, sometimes combining complex concepts. It is not trivial for a human to discard some of them without actually testing and comparing them.
- Most of the time produces a high-level recipe to evaluate the proposed ideas, missing details

Thanks!

Q&As

EAIRA EVAL Methodology :
<https://arxiv.org/abs/2502.2030>

9

