

- Context
- Why we did it
- Organizational challenges
- What we learn from it



Idea and lead organizer: Rick Stevens

DOE Labs experience exploring large AI models for science

Franck Cappello

Lead of AuroraGPT/Eval

Argonne National Laboratory

# Large Language Models (LLMs) Progress/4-5 years

Large Language Models (LLMs) have progressed drastically in the past 4-5 years (GPT3 released in 2020)

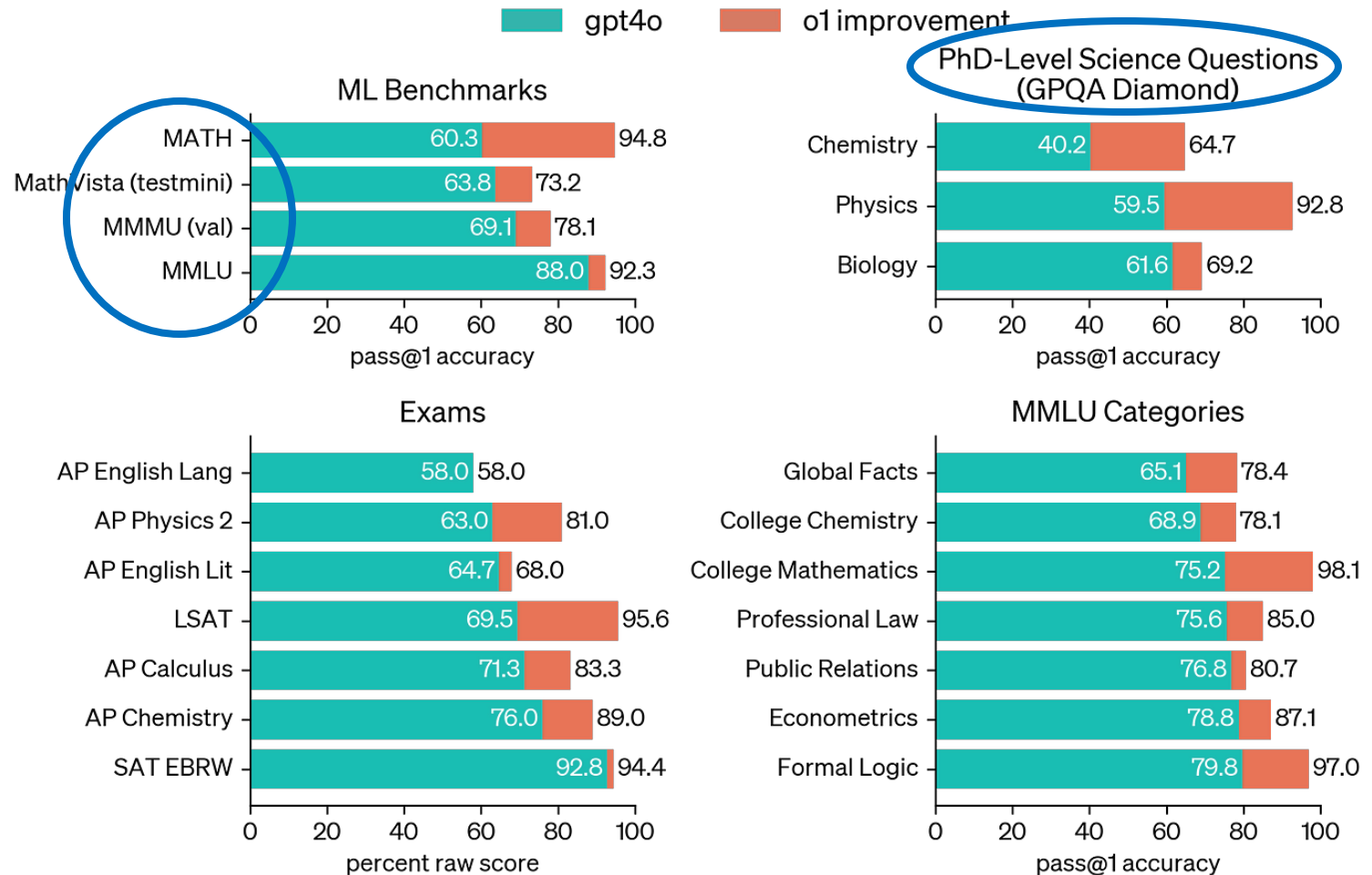
OpenAI's GPT4o (OpenAI 2024), Google's Gemini (Gemini 2024), and Anthropic's Claude (Anthropic 2024) are **excelling in text processing: summarization, information extraction, translation, and classification.**

Until recently (September 2024), Model performance (accuracy) progressed by increasing the size of the model and increasing the size of the training sets: **Trillions params/tokens**

On Sep. 12, 2024 OpenAI released **O1-preview: trained for reasoning**. Chain-of-thoughts + Reinforcement Learning during training. Internal chain-of-thoughts during inference.

→ **Greatly changed perception of what LLLs may be able to accomplish** in the near future.

Based on or adapted from classical test theory (CTT) in psychometrics



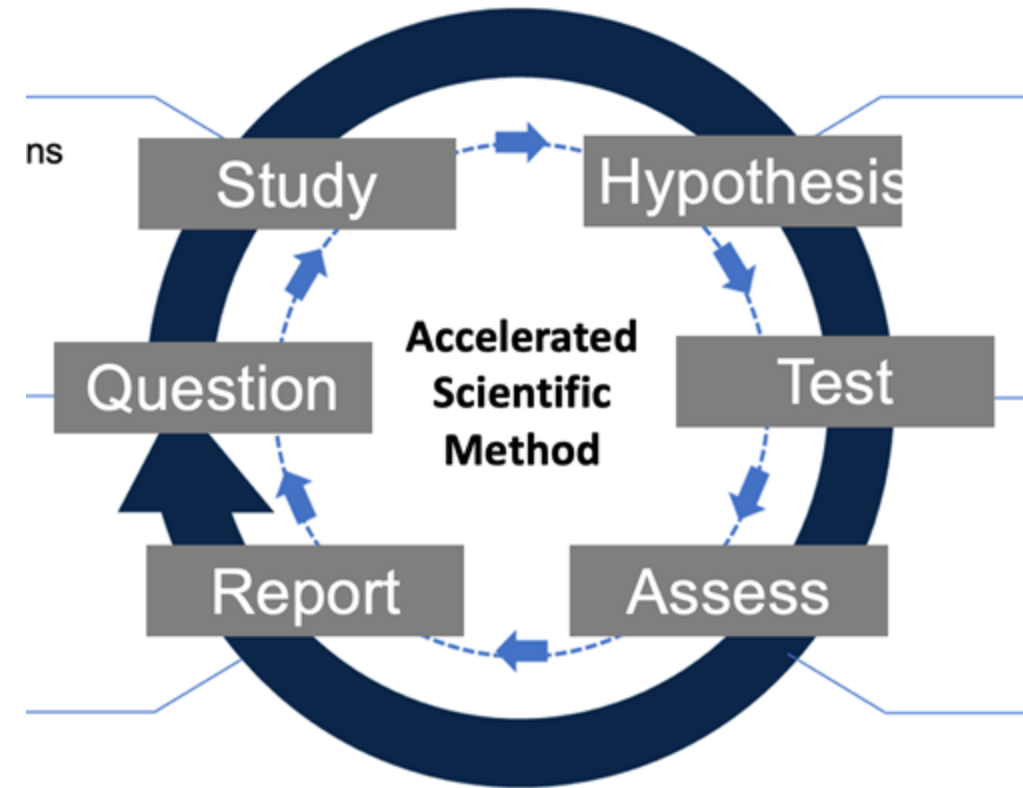
<https://sebastian-petrus.medium.com/openais-o1-mini-vs-o1-preview-a-comprehensive-comparison-b5d7b148dbda>

# LLMs/rLLMs as Research Assistants?

Scientists assessed LLMs on **specific tasks**:

- Predicting molecular properties
  - Uncovering genomic patterns
  - Solving mathematical problems
  - Creating and manipulating tools for simulations and analysis
  - Etc.
- Diversity and strength of skills and capabilities (Knowledge, Reasoning, Web search, Tools) **Suggest a new holistic approach** where LLMs/rLLMs are use as **scientific research assistants**

<https://doi.org/10.1038/s41524-022-00765-z>



# Three Main Challenges Before Broad Adoption

1) Encourage/Recommend Researcher to test LLMs for the different steps of the research circle **(Can LLMs accelerate research?)**

→ Many researchers are aware of these models but barely tried to use them for their research

2) Researchers need a way to **evaluate/compare the capabilities of LLMs in research context** for the different stages and tasks of the scientific research process **(which one is the best for what task?)**.

→ Need methods to evaluate and compare LLMs in research context

3) As with other research tools and techniques, **researchers will adopt LLMs only if they trust their results** **(Can I, Should I trust this response?)**.

→ Need a way to assess the correctness of the produced results, in order to develop confidence in their use in scientific context.



# The EAIRA Evaluation Methodology for the Evaluation of LLMs

End-to-End

New

New



## Proposed Methodology

In the Wild

Techniques	MCQ Benchmarks	Open Response Benchmarks	Lab Style Experiments	Field Style Experiments
Main Goal	Testing knowledge <b>breadth, basic reasoning</b>	Testing knowledge <b>depth, planning, reasoning</b>	<b>Realistic</b> testing	<b>Realistic trend</b> analysis and weakness diagnosis
Problem Type	<b>Predetermined</b> , Fixed Q&As with known solutions	<b>Predetermined</b> , Fixed Free-Response Problems with known solutions	<b>Individual Human</b> Defined Problems with <b>unknown</b> solutions	<b>Many Human</b> Defined Problems with <b>(un)known</b> solutions
Verification	<b>Automatic</b> response verification	<b>Automatic or Human</b> response verification	<b>Humans detailed</b> response analysis	Scalable <b>automatic</b> summary of <b>human response</b>
Examples	<b>Astro, Climate, AI4S</b> (multi-domain), <b>Existing Benchmarks</b>	<b>SciCode, ALDbench</b>	see "lab style experiments"	see "field style experiments"
Cross Cutting Aspects	← <b>Trust and Safety (ChemRisk), Uncertainty Quantification, Scalable Software Infrastructure (STAR)</b> →			

<https://arxiv.org/abs/2502.20309>  
or search EAIRA on Google

# Argonne Jam Sessions

*Argonne researcher participation and contribution on a voluntary basis.*

**First session: November 1, 2024**

## Argo/O1-Preview JAM Session

### Introducing LRM to researchers + training session

200 researchers from across Argonne spent two hours tackling a “staff-level” problem with Argo/O1-preview

- Asked researchers to **solve a research problem** (1h-1h30)
- **Used O1-preview Chat interface** (different “system” prompt compared to OpenAI O1-preview web version)
- 180 experiments in total
- **134 Experiments from various scientific domains** (after removing non non-scientific questions and experiments requiring only recall capabilities)
- Use the same **5 criteria and questions** as the ones used by LANL in their recent report
- Collected in addition, **scoring explanations and full conversations**
- Total of **19 questions/boxes**



**By the end of March 2025, All Argonne Directorates organized a JAM Session**



# ~~1000~~ 1500 Scientists AI JAM in 9 Labs Simultaneously (Feb.28, 2025)



*Researcher participation and contributions on a voluntary basis.*



# 1000 Scientists AI JAM Session: Goal and Rules of engagement



*Researcher participation and contribution on a voluntary basis.*

## Goals:

- Give Lab researchers an opportunity to test the best available LLMs
- Build a large corpus of interactions between researchers and AI models
  - Will help Labs understand how researchers will use reasoning models LLMs for Science →

### **How AI models may accelerate discoveries**

- Will help AI labs (OpenAI, Anthropic) to improve their model → to improve our research

## Rules:

- Explore advanced AI models on **challenging scientific problems**,
- Better understand the potential impact of AI reasoning models on **national security and science**,
- **In-person event** hosted at Argonne, Berkeley, Brookhaven, Idaho, Livermore, Los Alamos, Oak Ridge, Pacific Northwest, and Princeton Plasma Physics national laboratories. Scientists from other DOE labs are also participating,
- Explore models from OpenAI (**o1-pro, o1-deepresearch, o3-mini-high**) and Anthropic (**Claude 3.7 extended**),
- **OpenAI people in the rooms.**



# 1000 Scientists AI JAM Session: Rules of engagement (cont.)

*Researcher participation and contributions on a voluntary basis.*



## Rules (cont.):

- **Each participant brought 2-3 problems** from their scientific domain,
- Diverse range of realistic, representative tasks in scientific research and development,
- **Spend the day** using the latest capabilities of AI reasoning models,
- Participants can work in small teams to avoid getting stuck and to help observe or analyze the performance of the models,
- **All prompts, model responses and participant assessments are recorded --> the AI JAM Corpus,**
- The results will give an early estimate of how these tools could benefit the scientific community,
- The AI JAM Corpus is **shared with all participating DOE labs and AI labs (industry),**
- **Future sessions** will feature models from different AI companies (XAI, Nvidia, Google).

# 1000 Scientists AI JAM Session: Collection

*Researcher participation and contributions on a voluntary basis.*

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)

Prompts

Attachment

Model Response

Response assessment

# 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**





# What we learn from the 1000 Scientists AI JAM

- **Benchmarking is not enough** (we knew it, but real-life JAMs affirm it)
- Demonstrating AI models' capabilities to researchers is **important for adoption**
- **Organization was very challenging** (legal aspects)
- **Participants tested all aspects** of the research circle (except text processing...)
- The corpus represents the **largest existing collection of scientists/AI models' interactions**
- **Mining this corpus will help understand what researchers need and the gaps between the current situation and effective AI research assistants**

# First large scale collaboration with AI Industry

- **AI labs (OpenAI, Anthropic) are very interested** in scientific research (reasoning)
- What's our added value:
  - **1) We produce/have the scientific data AND we know how to transform it for injection.**
  - **2) We have the scientific knowledge/skills in many science domains**
  - **3) We will use LLMs for science only if we trust them.**

# Thanks!

## Q&As

