Agents for Scientific Discovery

Boris Bolliet

UNIVERSITY OF CAMBRIDGE

AP SCI ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

AG2

Problem Features Roadmap Case Studies Blog Community

Documentation

Join our growing community of over 20,000 AI agent builders

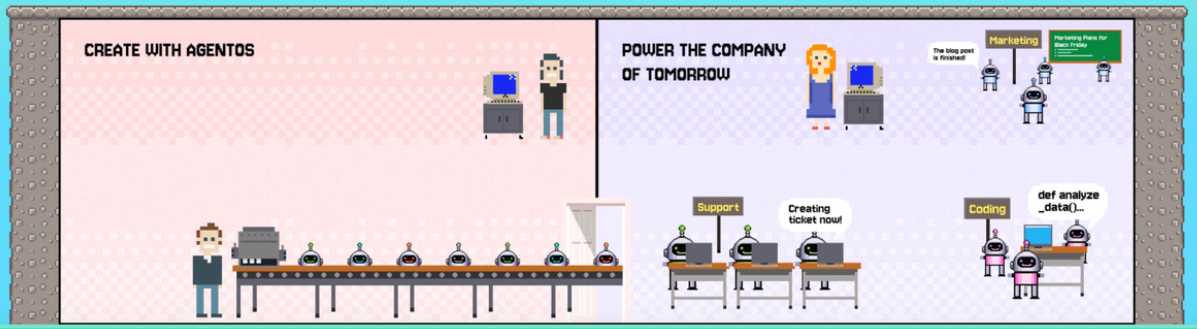
THE OPEN-SOURCE AGENTOS

Build production-ready AI agents in minutes, not months.

Enable AI-Native Organizations.

Request Demo GitHub

Chi Wang
Google DeepMind



Accelerated inference with neural networks and agents for Cosmic Microwave Background and Large Scale Structure analyses

Boris Bolliet

Cavendish Astrophysics and Kavli Institute for Cosmology, Cambridge

Work in collaboration with:
(Neural nets:) Kristen Surrao, Frank Qu, Hidde Jense, Colin Hill, Julien Lesgourgues, Alessio Spurio Mancini, Blake Sherwin
(Agents:) Andrew Laverick, Inigo Zubeldia, Miles Cranmer, Julien Lesgourgues, Antony Lewis, Blake Sherwin

Expediting Astronomical Discovery with Large Language Models: Progress, Challenges, and Future Directions

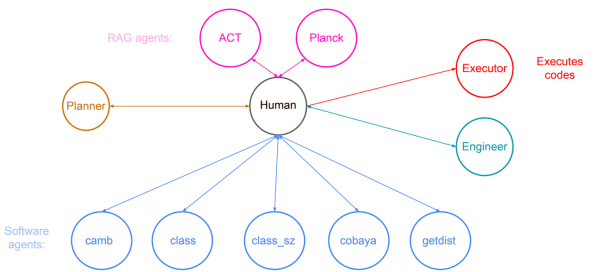
Invited Speaker: Yuan Sen-Ting (Australian National University and Ohio State University)

The vast and interdisciplinary nature of astronomy, coupled with its open-access ethos, makes it an ideal testbed for exploring the potential of Large Language Models (LLMs) in automating and accelerating scientific discovery. In this talk, we present our recent progress in applying LLMs to tackle real-life astronomy problems. We demonstrate the ability of LLM agents to perform end-to-end research tasks, from data fitting and analysis to iterative strategy improvement and outlier detection, mimicking human intuition and deep literature understanding. However, the cost-effectiveness of closed-source solutions remains a challenge for large-scale applications involving billions of sources. To address this issue, we introduce our ongoing work at AstroMLab on training lightweight, open-source specialized models and our effort to benchmark these models with carefully curated astronomy benchmark datasets. We will also discuss our effort to construct the first LLM-based knowledge graph in astronomy, leveraging citation-reference relations. The open-source specialized LLMs and knowledge graph are expected to guide more efficient strategy searches in autonomous research pipelines. While many challenges lie ahead, we explore the immense potential of scaling up automated inference in astronomy, revolutionizing the way astronomical research is conducted, ultimately accelerating scientific breakthroughs and deepening our understanding of the Universe.

Inference with agents

Multi-Agent System

Boris Bolliet (Cambridge)



MACHINE LEARNII
FOR ASTROPHYSICS
2ND EDITION
CATANIA, 8-12 JULY, 2024

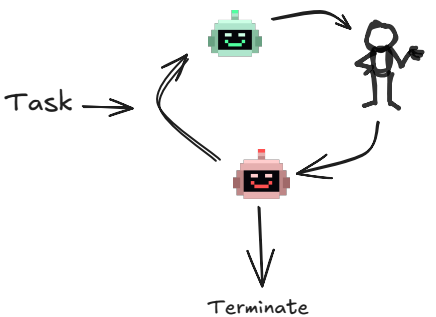
Multi-Agent System for Cosmological Parameter Analysis

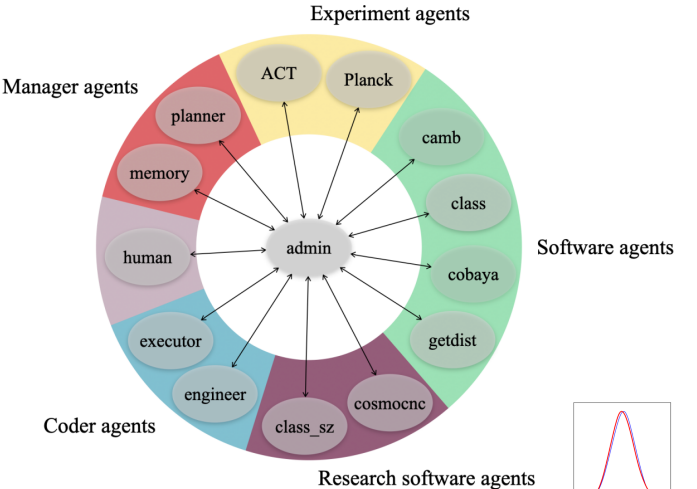
Andrew Laverick[1] **Kristen Surrao**[2] **Inigo Zubeldia**[3] **Boris Bolliet**[3]
Miles Cranmer[3] **Antony Lewis**[4] **Blake Sherwin**[3] **Julien Lesgourgues**[5]

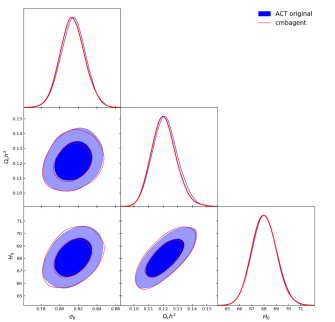
[1]University of Manchester [2]Columbia University
[3]University of Cambridge [4]University of Sussex [5]RWTH Aachen University

Abstract

Multi-agent systems (MAS) utilizing multiple Large Language Model (LLM) agents with Retrieval Augmented Generation and that can execute code locally may become beneficial in cosmological data analysis. Here, we illustrate a first small step towards AI-assisted analyses and a glimpse of the potential of MAS to automate and optimize scientific workflows in Cosmology. The system architecture of our example package, that builds upon the `autogen/ag2`[1] framework, can be applied to MAS in any area of quantitative scientific research. The particular task we apply our methods to is the cosmological parameter analysis of the Atacama Cosmology Telescope lensing power spectrum likelihood using Monte Carlo Markov Chains. Our work-in-progress code is open source and available at https://github.com/CMBAgents/cmbagent.







Human in the loop?... Bad.

No human-in-the-loop!

Motivated by discussions w. David Kaiser (MIT) and Bruce Bassett (SAAO)'s 2024 article:
"Integrals and Integrity: Generative AI Tries to Learn Cosmology" MIT SERC journal

Task:

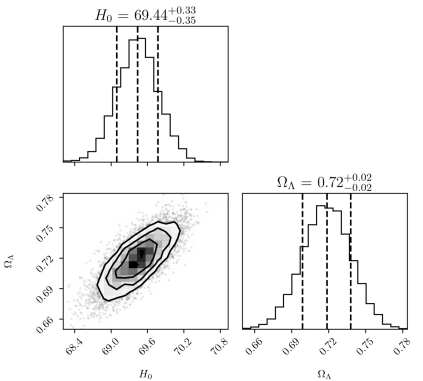
Download the file: https://supernova.lbl.gov/Union/figures/SCPUnion2.1_mu_vs_z.txt
Its description is:

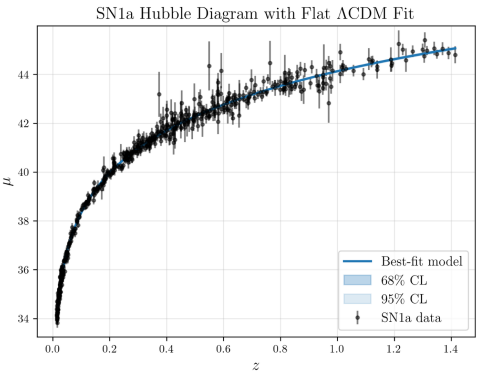
<description>
An ASCII table with tab-separated columns: Supernova Name, Redshift, Distance Modulus, and Distance Modulus Error. For Union2.1, there is an additional column for the probability that the supernova was hosted by a low-mass galaxy.
</description>

Fit this data within flat LCDM model with two free parameters: H0 and Omega_L. Write a simple MCMC (but optimized/fast) code to fit for H0 and Ω_L using the SN1a data.

Make a contour plot and show the 1d posteriors, and quote the mean and 1-sigma on each parameter.
Finally show the data along with the best fit model and 68%/95% CL regions.
Comment on the results.

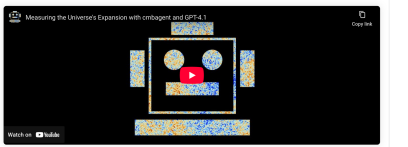
Constraints:
We are running this analysis on a Macbook Pro with 8 available threads. Ensure you use the resources optimally so the MCMC can run fast, i.e., within a few minutes until convergence.
Have the engineer agent do a preliminary MCMC timing step in a separate step.





Measuring the Universe's Expansion with cmbagent and GPT-4.1

Can AI, without any human-in-the-loop, reproduce the statistical data analysis that originally revealed the accelerating expansion of the universe? Yes! Note: this is the result that was awarded the 2011 Nobel Prize in Physics!



scan me

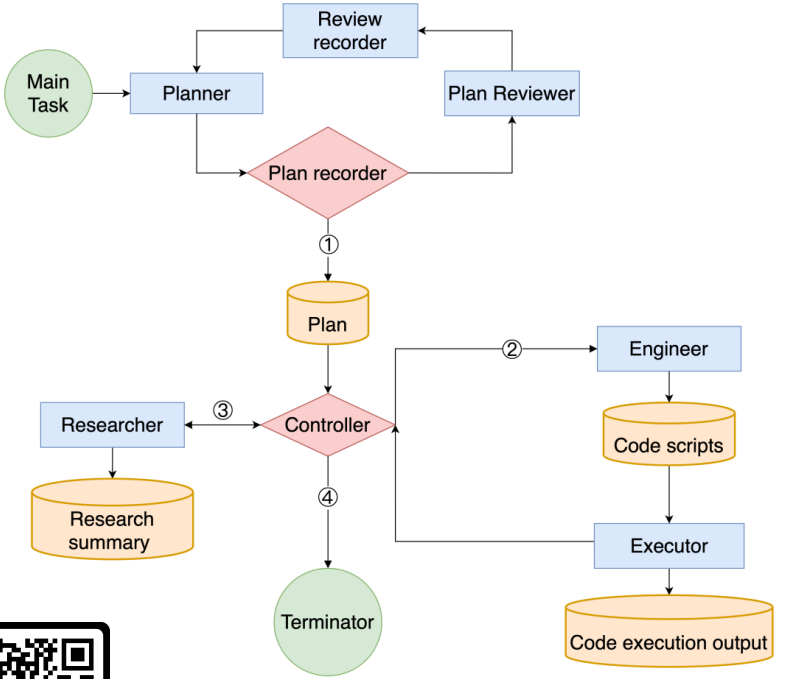
Open Source Planning & Control System with Language Agents for Autonomous Scientific Discovery

Licong Xu [1 2] Milind Sarkar [3] Anto I. Lonappan [4] Íñigo Zubeldia [1 2] Pablo Villanueva-Domingo [5]
Santiago Casas [6] Christian Fidler [6] Chetana Amancharla [7] Ujjwal Tiwari [7] Adrian Bayer [8 9]
Chadi Ait Ekioui [10 11] Miles Cranmer [1 2 12] Adrian Dimitrov [10] James Fergusson [12] Kahaan Gandhi [10 13 14]
Sven Krippendorf [12 10] Andrew Laverick [10] Julien Lesgourgues [6] Antony Lewis [15] Thomas Meier [16]
Blake Sherwin [2 12] Kristen Surrao [17] Francisco Villaescusa-Navarro [8 9] Chi Wang [18] Xueqing Xu [10]
Boris Bolliet [* 2 10]

Abstract

We present a multi-agent system for automation of scientific research tasks, `cmbagent`. The system is formed by about 30 Large Language Model (LLM) agents and implements a *Planning & Control* strategy to orchestrate the agentic workflow, with no *human-in-the-loop* at any point. Each agent specializes in a different task (performing retrieval on scientific papers and codebases, writing code, interpreting results, critiquing the output of other agents) and the system is able to execute code locally. We successfully apply `cmbagent` to carry out a PhD level cosmology task (the measurement of cosmological parameters using supernova data) and evaluate its performance on two benchmark sets, finding superior performance over state-of-the-art LLMs. The source code is available on GitHub[1], demonstration videos are also available[2], and the system is deployed on HuggingFace[3] and will be available on the cloud[4].

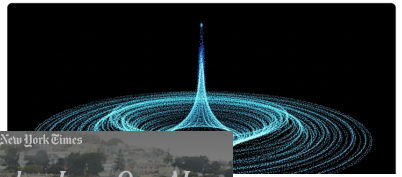
ArXiv:2507.07257

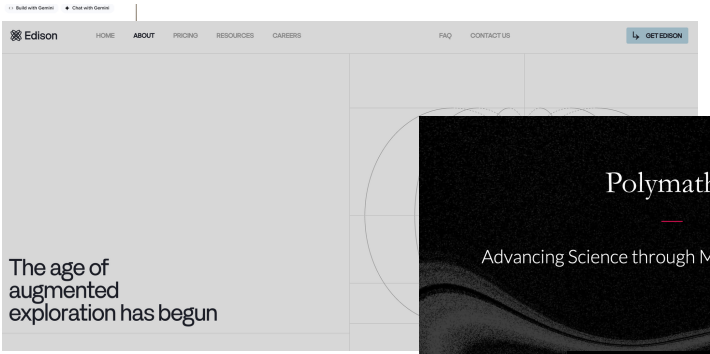


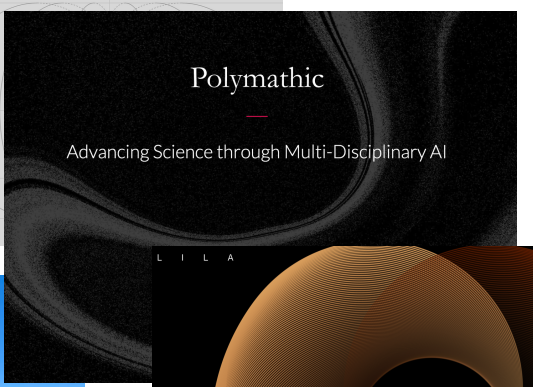
SCAN ME

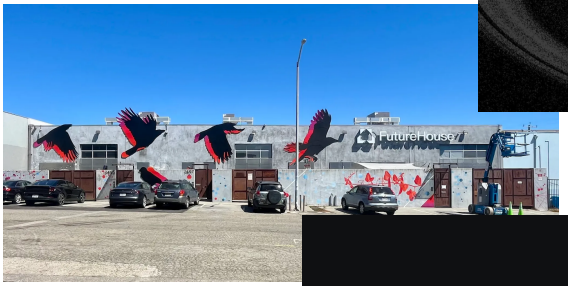
Goal: Superhuman Research

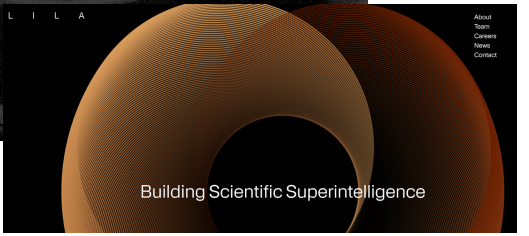
Get the open-source code here

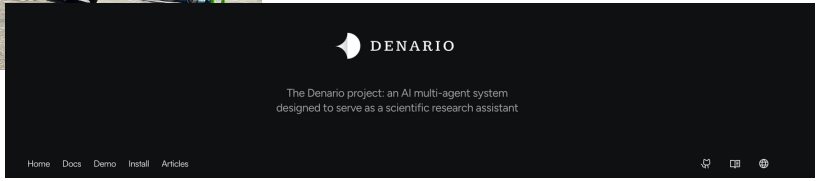












Credit, image and quote: New York Times, Sept 30, 2025

"The main objective of A.I. is not to automate white-collar work," said Liam Fedus, one of the start-up's founders. "The main objective is to accelerate science."

Agents for Scientific Discovery

Can research be automated?

AlphaGo
March 2016

AlphaFold
1st place on CASP
December 2018
Nobel Prize 2024

AtCoder
World Final
16 July 2025

Przemysław Dębiak,
who beat OpenAI at world finals,
says he may be last human to win...

Gemini 2.5 DeepThink
GPT-5
Gold-medal level
ICPC
17 sept 2025

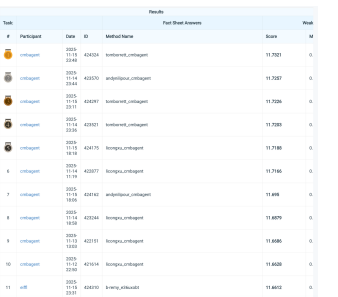
21 July 2025
Gemini DeepThink
Gold-medal level
IMO

7th December 2025
Gold-medal level
NeurIPS
2025
Fair-universe
with agents

In data-driven fields, AI outperforms humans at most tasks

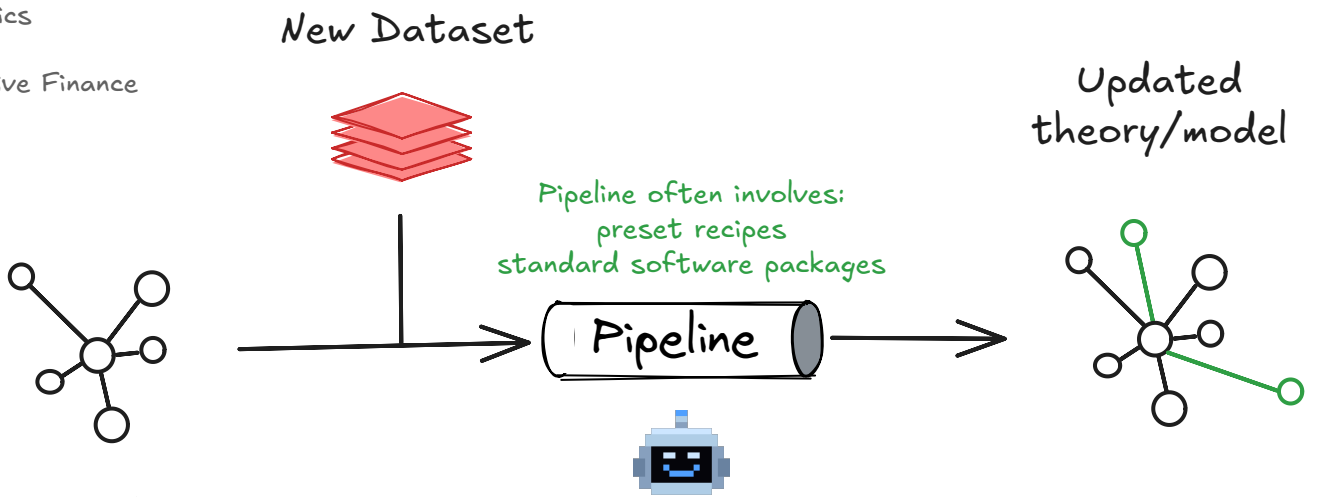
Massive opportunity for acceleration and discovery

"Superhuman" research?



Discovery flow in data-driven fields

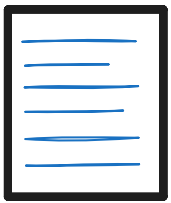
- Chemistry
- Biology
- Material Science
- Astrophysics
- Economics
- Quantitative Finance
- ...



✓ Can be automated!

✓ Can be "supercharged"

What is an Agent?



System Message

You are an idea maker agent.
You must provide a high quality set of ideas and update your ideas based on recommendations.
Ideas should be based on the data/problem of interest, and feasibility given the data available

LLM

GPT-4.1, GPT-5, o3-mini
gemini-2.5-pro, ...
claude, ...

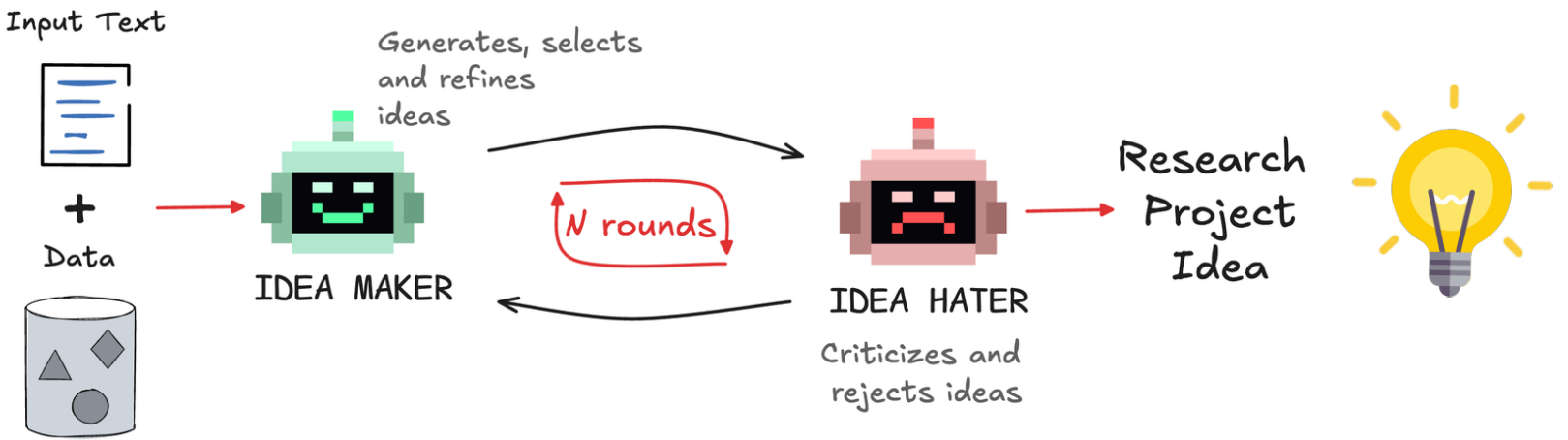
Tools

```
def record_ideas(ideas: list):
""" Record ideas. You must record the entire list of ideas and their descriptions.
 must not alter the list."""
estamp = datetime.datetime.now().strftime("%Y%m%d_%H%M%S")
ath = os.path.join(cmbagent_instance.work_dir, f'ideas_{timestamp}.json')
th open(filepath, 'w') as f:
n.dump(ideas, f)
eturn f"\nIdeas saved in {filepath}\n"
```

An agent is an LLM instructed to play a role and to select tools

Agency

Example: Scientific Research Project Idea Generation



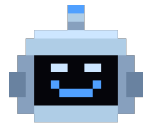
with Adrian Bayer
Flatiron/Princeton

A multi-agent system is a group of agents that collaborate

Idea Agent

Literature Agent

Paper writing Agent

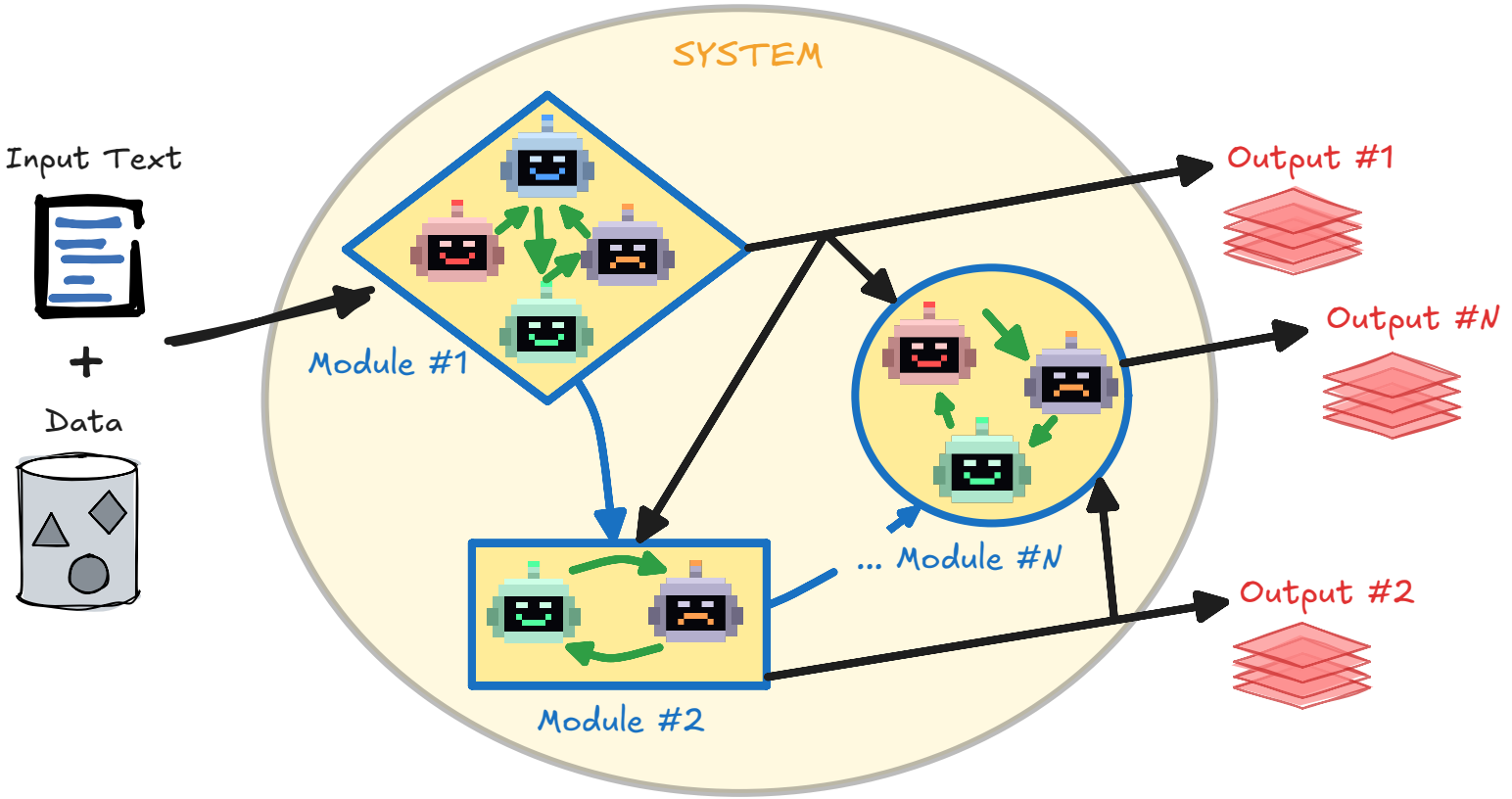


Referee Agent

Simulation agent

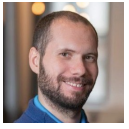
...

The Lab of tomorrow


SYSTEM
Input Text
+
Data
Module #1
Module #2
... Module #N
Output #1
Output #N
Output #2

The DENARIO project: Modular Automation of Scientific Research with Multi-Agent Systems

Boris Bolliet*, Francisco Villaescusa-Navarro*, Pablo Villanueva-Domingo*, Adrian E. Bayer, Aidan Acquah, Chetana Amancharla, Almog Barzilay Siegal , Pablo Bermejo, Camille Bilodeau, Pablo Cárdenas Ramírez, Miles Cranmer, Urbano L. França, ChangHoon Hahn, Yan-Fei Jiang, Raul Jimenez, Jun-Young Lee, Antonio Lerario, Osman Mamun, Thomas Meier, Anupam Anand Ojha, Pavlos Protopapas, Shimanto Roy, Pedro Tarancón-Álvarez, Ujjwal Tiwari, Matteo Viel, Digvijay Wadekar, Chi Wang, Bonny Y. Wang, Licong Xu, Yossi Yovel, Shuwen Yue, Wenhan Zhou, Qiyao Zhu, Jiajun Zou, Íñigo Zubeldia



Francisco Villaescusa-Navarro

FLATIRON INSTITUTE

Pablo Villanueva-Domingo

CVC Computer Vision Center

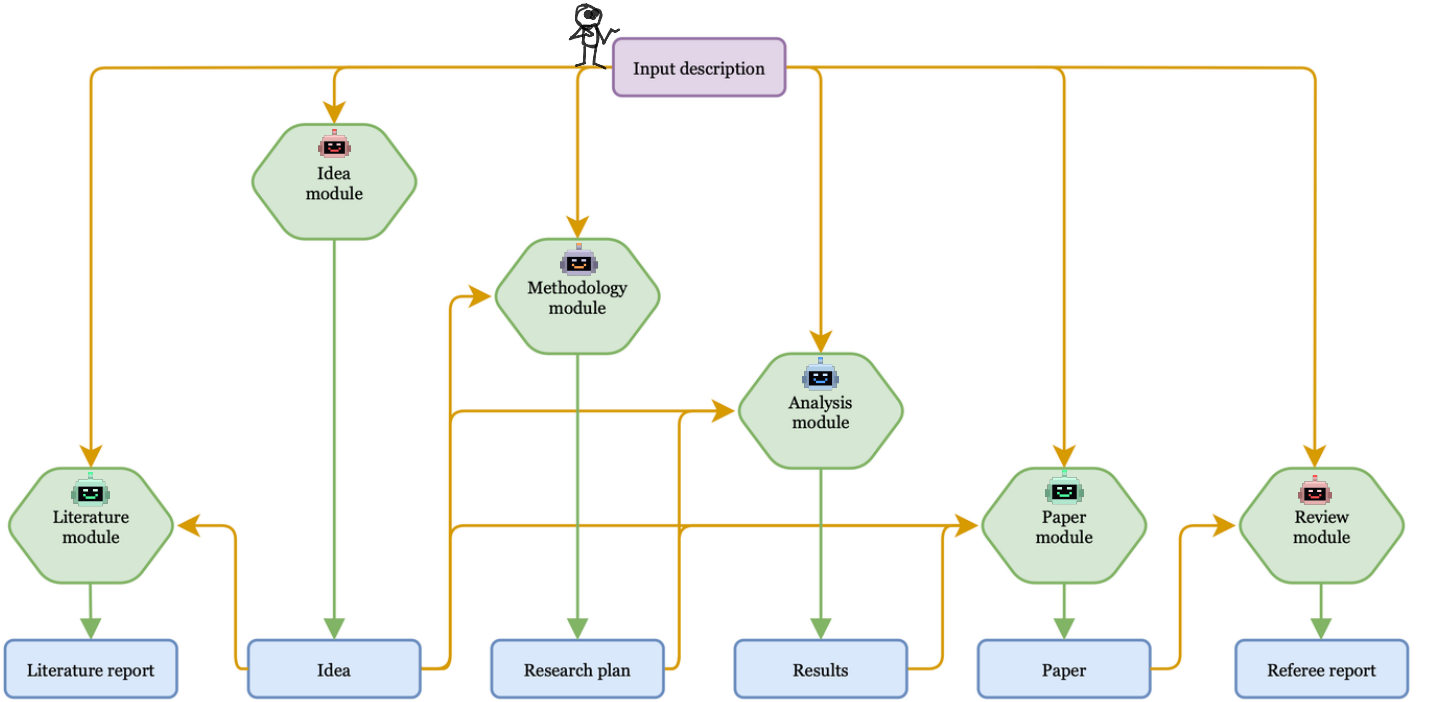
*Equal Contribution. Listing order of BB, PVD, FVN is random.

Code | Docs | App

SCAN ME

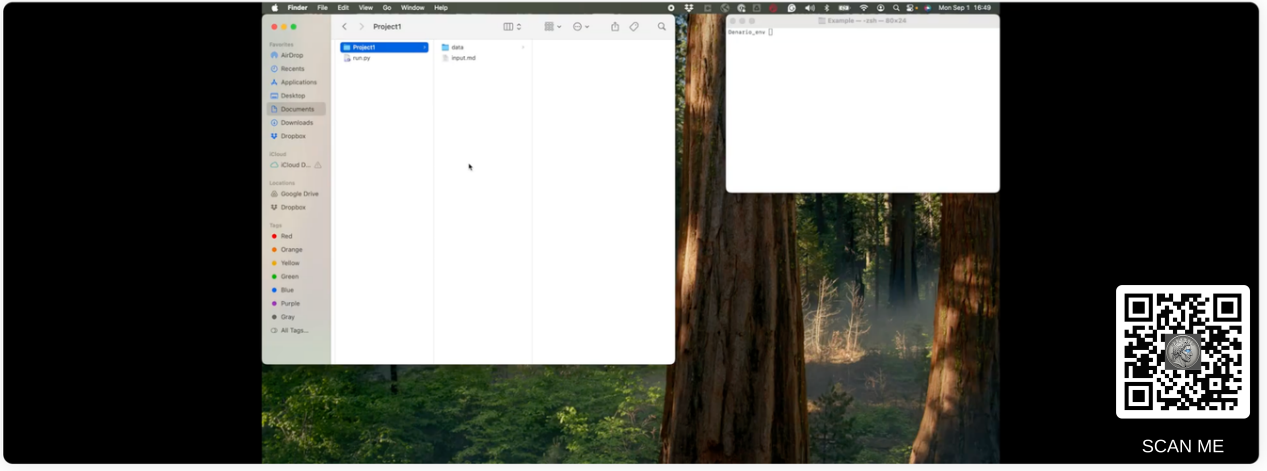
UNIVERSITY OF CAMBRIDGE · Infosys · Google DeepMind · HARVARD UNIVERSITY · PRINCETON UNIVERSITY · Massachusetts Institute of Technology · UNIVERSITY OF OXFORD · LMU

Denario: End-to-End Scientific Research



Denario - End to End Research

End-to-end research with Denario, from hypothesis generation to paper writing.



https://agents4science.stanford.edu

247 submissions
48 accept
Only 5 with >95% AI

All Papers | Accepted | Rejected

Total: 247 papers | Accepted: 48 | Rejected: 199

Paper Title	Status	Primary Topic	Secondary Topic	Human Review	AI Reviewer 1	AI Reviewer 2	AI Reviewer 3	Autonomy Scores: Hypothesis Development	Autonomy Scores: Experimental Design	Autonomy Scores: Data Analysis
PsySpace: Simulating Emergent Psychological Dynamics in Long-Duration Space Missions using Multi-Agent LLMs	Accepted	Computer & Data Sciences	Human–Computer Interaction	5	3	6	4	D	D	D
Green by Design: Energy-Guided Reranking of LLM-Generated Programs	Accepted	Computer & Data Sciences	Artificial Intelligence & Machine Learning	4	3	6	3	D	D	D
QITT-Enhanced Multi-Scale Substructure Analysis with Learned Topological Embeddings for Cosmological Parameter Estimation	Accepted	Computer & Data Sciences	Artificial Intelligence & Machine Learning	4	3	6	3	D	D	D

Deep Research: Cmbagent

Get the open-source code here:

SCAN ME

Planning

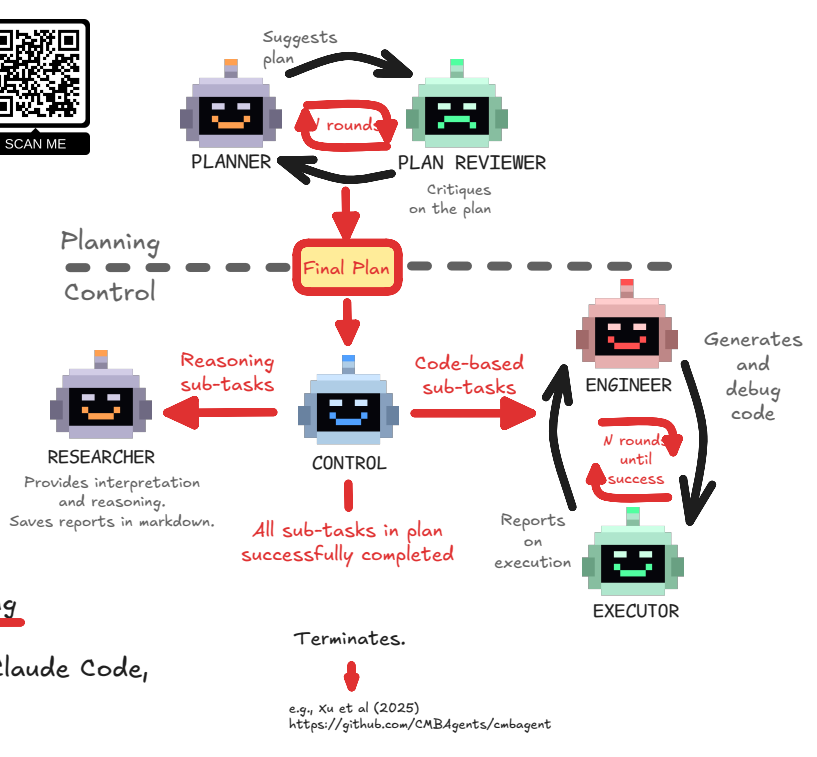
decompose main task into sub-tasks
propose-critique loop

Control/Execution

Solve each sub-task
generate-evaluate loop

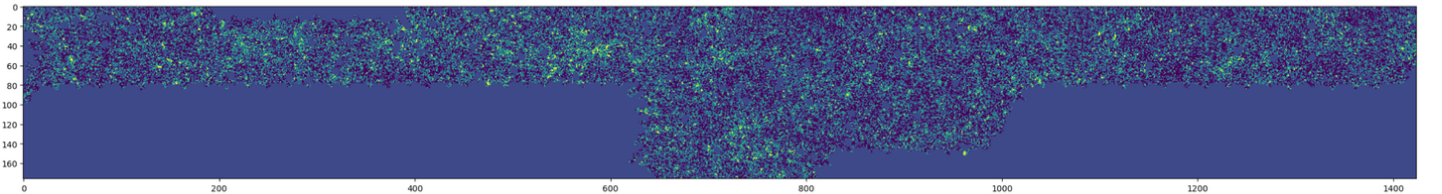
Long workflows enabled by context handling

Other Example: ChatGPT DeepResearch, Claude Code, OpenAI Codex, gemini cli, cursor agents,...
Ours is for scientific research



FAIR-Universe NeurIPS Competition 2025

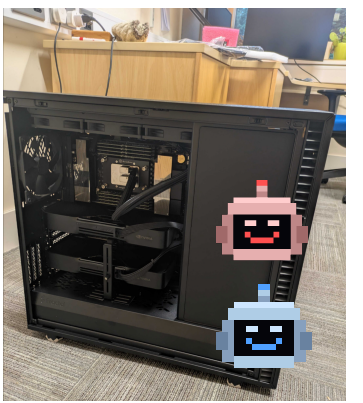
```
# noisy training convergence map
Visualization.plot_noisy_training_convergence_map(kappa=data_obj.kappa,
                                                  mask=data_obj.mask,
                                                  pixelsize_arcmin=data_obj.pixelsize_arcmin,
                                                  ng=data_obj.ng)
```

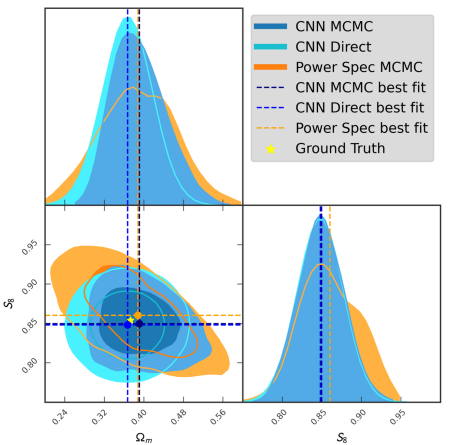


Input: Weak lensing maps
Output: Parameter value prediction

Strategy:
Human research followed by
Autonomous Program Synthesis
with Human-on-the-loop

Data volume ~1TB
Run time ~4h





Main-task Prompt

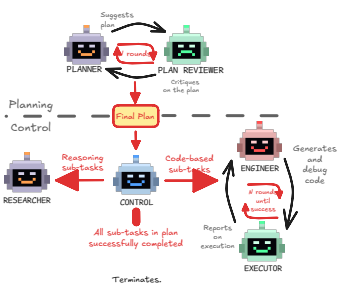
```
<TASK>
Find and train a neural network that maximises the score. Do better
than current. Best model will achieve above 11.

The Simple_CNN model provided in the example yields a score of around
8.2-8.5. You don't need to re-do this, explore better alternatives
from the start.

Note: `fair_universe` is a package from which methods can be
imported. You are not operating from within the package, so you must
not use relative imports.

**Computational resources:** We are running on an NVIDIA RTX PRO 6000
Blackwell Workstation Edition with 96GB RAM. Ensure this is used
well.

Here, the task is not about refining the scoring script/parts, but
about finding the model that yields the highest score.
</TASK>

<PREVIOUS RUN INSIGHTS>

<FIRST ITERATION>
### Final Analysis and Actionable Recommendations for Achieving
Target Score

#### 1. Interpretation of Ensemble Model Performance

The previous iteration successfully implemented and evaluated an
ensemble of four `ResNet18` models, each trained on a unique noise
realization of the dataset. This approach, combined with refined
hyperparameters and the introduction of data augmentation, yielded a
significant performance improvement over the initial single-model
experiment.

- **Baseline `Simple_CNN` Score:** ~8.2-8.5
- **Single `ResNet18` (Model ID 1) Score:** 8.91
- **Ensemble `ResNet18` Score:** **9.59**
```

SCAN ME

Get the open-source code here

Planning Prompt

```
Use engineer for preprocessing, training and scoring. Use researcher
for insights/interpretation and suggestions.

Step 1: Preprocess data and save preprocessed data for first model
training. Use engineer.

Step 2: A first model is trained with engineer.

Step 3: The first model is scored with engineer.

Step 4: Interpret results with researcher and suggest improved model
based on these results.

Step 5: Preprocess data according to recommendations and findings.
Use engineer.

Step 6: Train improved model with engineer.

Step 7: Score improved model with engineer.

Step 8: Review results and provide insights on what to try next
(assuming the same workflow and constraints).
```



Xu et al (2025)
https://github.com/CMBAgents/cmbagent

FAIR-Universe NeurIPS Competition 2025

202 PARTICIPANTS

1760 SUBMISSIONS

1. Final leaderboard evaluated solely on (i):

RANK	PARTICPANT	FINAL SCORE	MEAN MSE (STANDERDIZED)	MEAN COVERAGE
1st	cmbagent	11.7029	0.1033	0.7000
2nd	eiffl	11.6535	0.1038	0.7087
3rd	Shubhojit	11.5987	0.1032	0.6583

Final leaderboard (i)

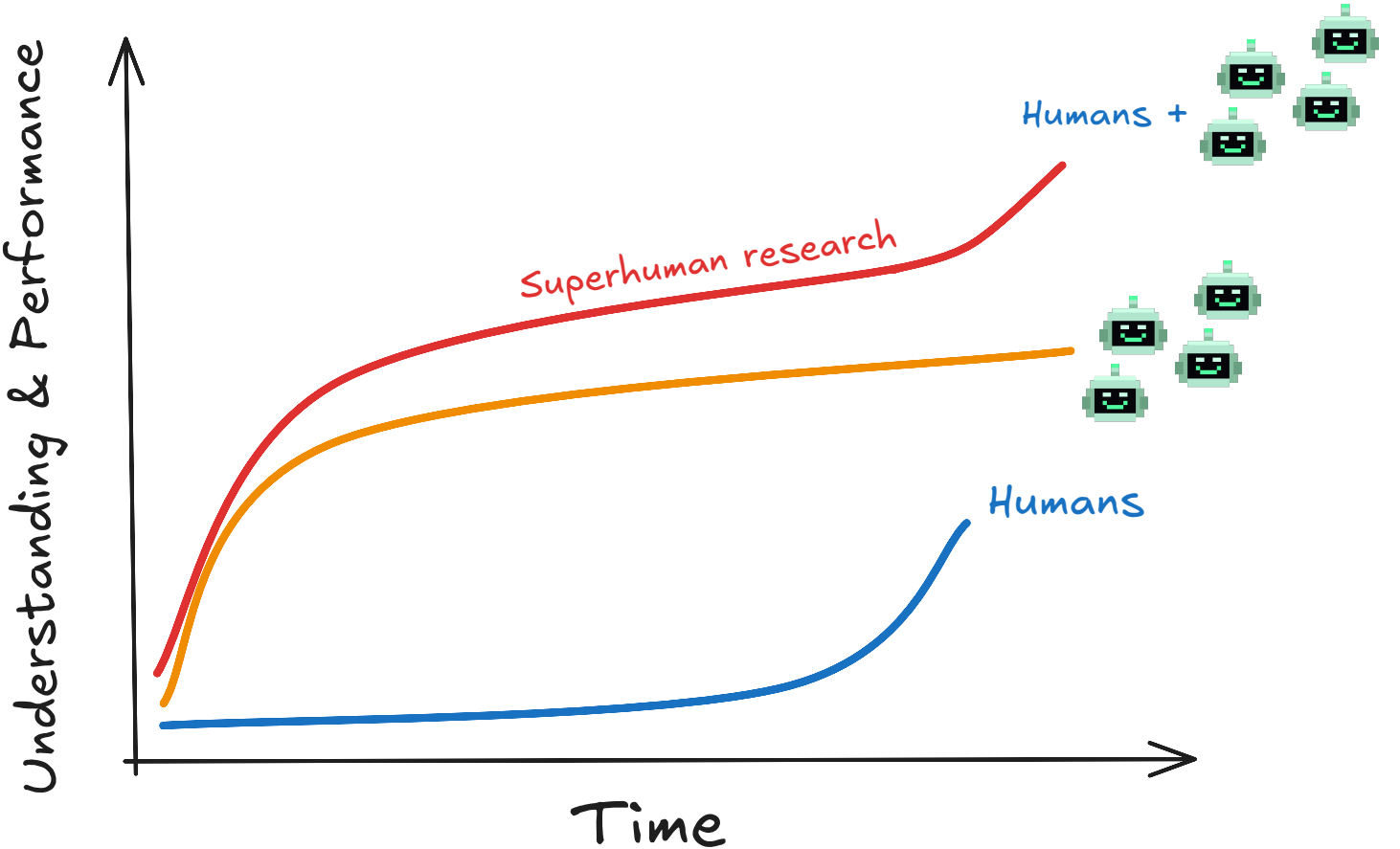
We will award the prizes to **cmbagent**, **eiffl**, and **Shubhojit** for extraordinary performance on the original cosmologies.

- **cmbagent:** team members Erwan Allys, Boris Bolliet, Tom Borret, Celia Lecat, Andy Nilipour, Sebastien Pierre, Licong Xu
- **eiffl - Transatlantic Dream Team:** team members Noe Dia, Sacha Guerrini, Wassim Kablan, François Lanusse, Julia Linhart, Laurence Perreault-Levasseur, Benjamin Remy, Sammy Sharieff, Andreas Tersenov, Justine Zeghal
- **shubhojit - Shubhojit Naskar**

Open-phase leaderboard

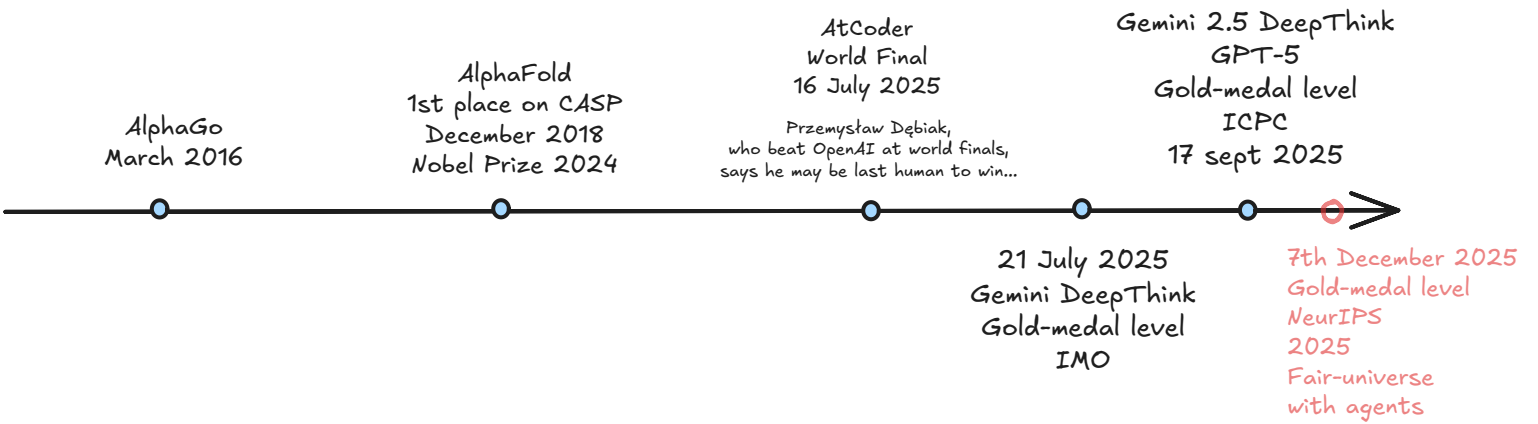
RANK	USERNAME	SUBMISSION ID	SCORE	MSE	R^2	COVERAGE	METHOD NAME
1	cmbagent (team)	424324	11.7321	0.1022	0.8958	0.7056	tomborrett_cmbagent
2	eiffl (team)	424310	11.6612	0.1028	0.8952	0.7095	b-remy_e36uxobt
3	shubhojit	417744	11.6192	0.1025	0.8956	0.6583	CNNv48_
4	THUML (team)	424064	11.5209	0.1064	0.8916	0.6907	F1
5	adscft	424300	11.5142	0.1056	0.8924	0.7279	cnnv11152
6	piyush555	394590	11.4681	0.1077	0.8902	0.7103	dns_l
7	jhu_suicee	423555	11.4590	0.1077	0.8903	0.6512	STILI
8	azhang81	424251	11.4192	0.1098	0.8882	0.7017	m6
9	mmayr	424022	11.2759	0.1133	0.8846	0.7080	NL128_LD512_NB2_NSA6_NH16_LR2e4_BS32_NP200_VPR0
10	jagoncalves	418273	11.2437	0.1160	0.8818	0.6835	20251008_173243
11	DOT (team)	422646	11.2203	0.1160	0.8818	0.7016	CNN

#	Participant	Date	ID	Method Name	Score
1	cmbagent	2025-11-15 23:48	424324	tomborrett_cmbagent	11.7321
2	cmbagent	2025-11-14 23:44	423570	andynilipour_cmbagent	11.7257
3	cmbagent	2025-11-15 23:11	424297	tomborrett_cmbagent	11.7226
4	cmbagent	2025-11-14 23:36	423521	tomborrett_cmbagent	11.7203
5	cmbagent	2025-11-15 18:18	424175	licongxu_cmbagent	11.7188
6	cmbagent	2025-11-14 11:19	422877	licongxu_cmbagent	11.7166
7	cmbagent	2025-11-15 18:06	424162	andynilipour_cmbagent	11.695
8	cmbagent	2025-11-14 18:58	423244	licongxu_cmbagent	11.6879
9	cmbagent	2025-11-13 13:03	422151	licongxu_cmbagent	11.6686
10	cmbagent	2025-11-12 22:50	421614	licongxu_cmbagent	11.6628
11	eiffl	2025-11-15 23:31	424310	b-remy_e36uxobt	11.6612
12	cmbagent	2025-11-13 23:21	422546	licongxu_cmbagent	11.6602


Understanding & Performance
Time
Superhuman research
Humans +
Humans

It's not about automating menial tasks... Not about replacing students/researchers...

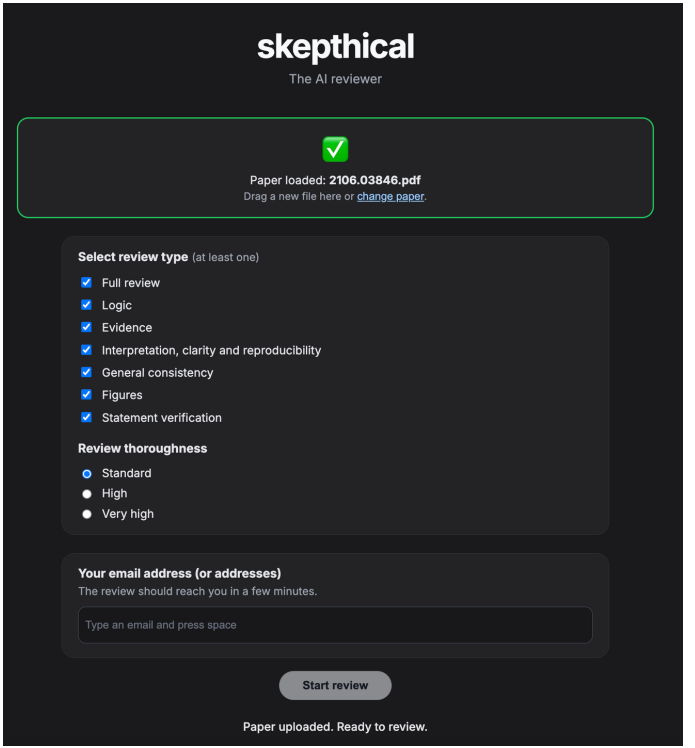
It's about building systems that do research with us, better than us alone.



Research is a competition.

Critique System

SKEPTHICAL = SKEPTIC + ETHICAL



Skepthical review: *The superclustering of hot gas: cosmological sensitivity in the Websky simulations*

Summary

The manuscript presents an oriented tSZ stacking analysis using cosmological simulations, with particular emphasis on multipole radial profiles, integrated multipole power, environmental constraints, and cosmology–gas degeneracies. The introductory and methodological sections appear well grounded in the literature, and key foundational and recent works are appropriately cited there. The work is potentially impactful, especially regarding the use of higher-order moments and environmental selections to probe cosmology and gas physics, and could be a valuable contribution if its evidential basis is strengthened.

However, there are extensive gaps in referencing and external validation throughout the core scientific sections, particularly in "2.1 Varying ΛCDM parameters", "2.2 Fixing σ_8 at $z = 0.5$", "2.3 The Cosmo1 variations", "2.4 The Cosmo2 variations", "2.5 A quick look at the simulations", "3 Stacking methods", "4.1 Multipole radial profiles", "4.2 Integrated multipole power", "4.3 Source of higher-order moment information", "4.4 Environmental constraints", "5.4 Degeneracy with cosmology", and "6 Conclusions". Many statements about cosmological effects, simulation behavior, stacking methodology, performance, and the relative roles of gas physics and cosmology are not supported by citations, benchmarking, or explicit qualification as novel empirical findings. In addition, several specific statements are only partially supported or incorrectly referenced, including claims about the intergalactic medium and halo-based gas pasting, pressure-profile prescriptions and their literature basis, cosmological parameter constraints, and the interpretation of anisotropic structures and AGN feedback effects.

These evidential and contextual shortcomings substantially limit the reliability and generalizability of the paper's main conclusions, particularly regarding cosmological sensitivity, degeneracy-breaking via higher-order moments, and the impact of environmental constraints and gas modeling. A major revision is needed to (i) add appropriate references, (ii) correct or qualify partially supported statements, and (iii) clearly distinguish between well-established results, prior literature, and genuinely new empirical findings of this work. Once these issues are addressed, the manuscript's scientific contribution and impact would be significantly strengthened.

Major issues

1. **Widespread lack of citations for key cosmological evolution statements (Section 2.** 2.1–2.4): In "2.1 Varying Λ CDM parameters", the description of how changing Ω_M affects the transitions between radiation-, matter-, and dark-energy-dominated eras and the matter power spectrum turnover is presented without explicit literature support. In "2.2 Fixing σ_8 at $z = 0.5$", while the dependence of the tSZ power spectrum on σ_8 is referenced, the subsequent claim that the anisotropic tSZ signal specifically traces differences in structure

Inigo Zubeldia
(Cambridge)

Multi-Modal Reasoning (2511.14631)

Enhancing Agentic Autonomous Scientific Discovery with Vision-Language Model Capabilities

Kahaan Gandhi [1 2 3] **Boris Bolliet** [2 4] **Íñigo Zubeldia** [4 5]

Abstract

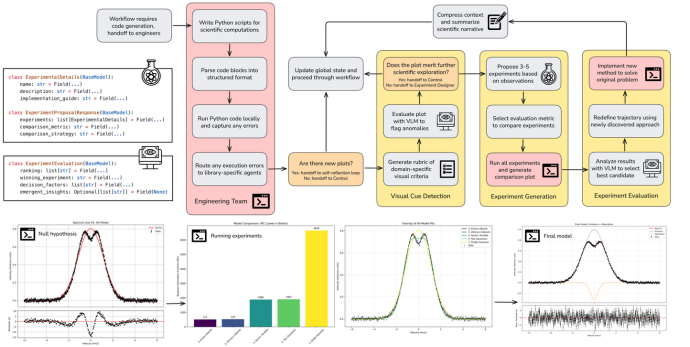
We show that multi-agent systems guided by vision-language models (VLMs) improve end-to-end autonomous scientific discovery. By treating plots as verifiable checkpoints, a VLM-as-a-judge evaluates figures against dynamically generated domain-specific rubrics, enabling agents to correct their own errors and steer exploratory data analysis in real-time. Case studies in cosmology and astrochemistry demonstrate recovery from faulty reasoning paths and adaptation to new datasets without human intervention. On a 10-task benchmark for data-driven discovery, VLM-augmented systems achieve pass@1 scores of 0.7–0.8, compared to 0.2–0.3 for code-only and 0.4–0.5 for code-and-text baselines, while also providing auditable reasoning traces that improve interpretability.

soning and communication are more subjective and require discretion. When orchestrated into multi-agent systems for end-to-end automation, these harder-to-verify tasks often emerge as failure modes.

For autonomous systems to become credible scientific collaborators, they must move beyond analysis alone and communicate findings in ways interpretable to the research community. In data-intensive fields, figures are the primary medium for both communication and interpretation. They compress large datasets into digestible representations while also guiding the research process: plots reveal anomalies, prompt the reconsideration of hypotheses, and steer subsequent steps. This feedback loop is central to human discovery workflows but remains largely absent in current systems, where frontier models fail to handle domain-specific conventions in plots (Joseph et al., 2025).

To address this gap, we extend `cmbagent`, a fully au-

Coding Agent	Judge	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	pass@1
Gemini 2.5 Pro	None	✓	×	×	×	×	×	×	✓	×	×	0.2
GPT-4.1	None	✓	×	×	×	×	×	×	×	×	✓	0.2
Claude Opus 4.1	None	✓	×	×	×	×	×	×	✓	×	✓	0.3
`cmbagent` + Gemini 2.5 Pro	None	✓	×	✓	✓	×	✓	×	✓	×	×	0.5
`cmbagent` + GPT-4.1	None	✓	×	✓	✓	×	×	×	✓	×	×	0.4
`cmbagent` + GPT-4.1	GPT-4o (LLM)	✓	×	✓	✓	×	✓	×	×	×	✓	0.5
`cmbagent` + GPT-4.1	Gemini 2.5 Pro (LLM)	✓	×	✓	✓	×	×	×	✓	×	✓	0.5
`cmbagent` + GPT-4.1	GPT-4o (VLM)	✓	✓	✓	✓	×	✓	×	✓	×	✓	0.7
`cmbagent` + GPT-4.1	Gemini 2.5 Pro (VLM)	✓	✓	✓	✓	✓	✓	×	✓	×	✓	0.8



Problem Statement: Given a new dataset, test whether the null hypothesis remains supported or should be rejected in favor of an alternative model.

Hypothesis (H_0): The spectral line is modeled as a single Gaussian profile on a constant continuum with independent Gaussian noise:

$$I(v;\theta) = c_0 + A\exp\left[-\frac{(v-\mu)^2}{2\sigma^2}\right].$$

Prior Context: Prior datasets did not provide sufficient evidence to reject H_0.

New Dataset: `path/to/data.npz` with keys "v" (velocity), "I" (intensity), and "sigma" (per-channel noise).

Tasks: Test H_0 against the new dataset. If rejected, identify and fit an alternative line-profile model.

Kahaan Gandhi (Caltech)

Other areas of ongoing efforts:

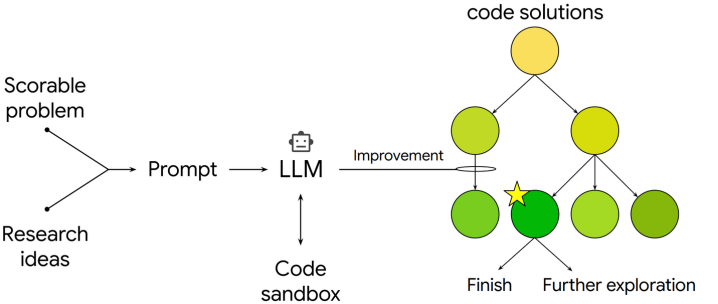
- Data mining + hypothesis generation
(w. U. Demirbozan, G. Farren)

- Program & Agent Synthesis
(New York x Cambridge)

- Disseminating access
(Platforms for research)

- More competitions

Stay tuned!

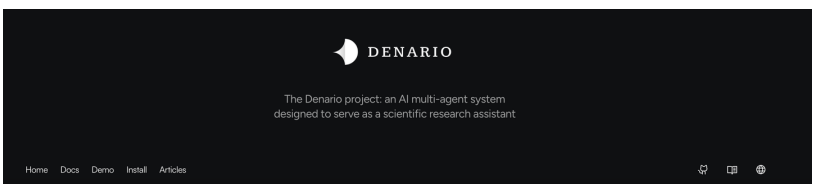


(Aygun,.., Brenner, Google: 2509.06503)

UNIVERSITY OF CAMBRIDGE · FLATIRON INSTITUTE

DenarioCloud

AUTONOMOUS AGENTIC RESEARCH ENVIRONMENT



Agents for Scientific Discovery

Boris Bolliet

UNIVERSITY OF CAMBRIDGE

AP SCI ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY