FOSS India

DiscoveryBench: Data-Driven Scientific Discovery with LLM Agents

Harshit Surana

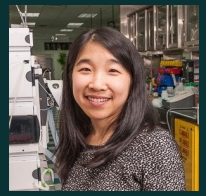
Allen Institute for Artificial Intelligence & OpenLocus












Carnegie Mellon University

Seattle Hub for SYNTHETIC BIOLOGY

ALLEN INSTITUTE for BRAIN SCIENCE

THE PAUL G. ALLEN FRONTIERS GROUP

OpenLocus

Ai2

Structure & Some Requests

1. Scientific Discovery with LLMs
2. DiscoveryBench/ Data Driven Discovery
3. Best Practices in Building AI at Scale

- Important to have interactivity
- Feel free to interrupt for quick clarifications.
- Will stop at the end of each section for **QnA**
- If audio issues arise, **please drop messages** in the chat.

Scientific Discovery with LLMs

Ai2

“Superintelligent tools could massively accelerate scientific discovery and innovation well beyond what we are capable of doing on our own, and in turn massively increase abundance and prosperity.”

- *Sam Altman, OpenAI*

NOBELPRISET I KEMI 2024
THE NOBEL PRIZE IN CHEMISTRY 2024

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

David Baker
University of Washington
USA

”för datorbaserad proteindesign”

“for computational protein design”

Demis Hassabis
Google DeepMind
United Kingdom

John M. Jumper
Google DeepMind
United Kingdom

”för proteinstrukturprediktion”

“for protein structure prediction”

#NobelPrize

THE NOBEL PRIZE

Broad Areas of Scientific Discovery where LLMs are Helping!

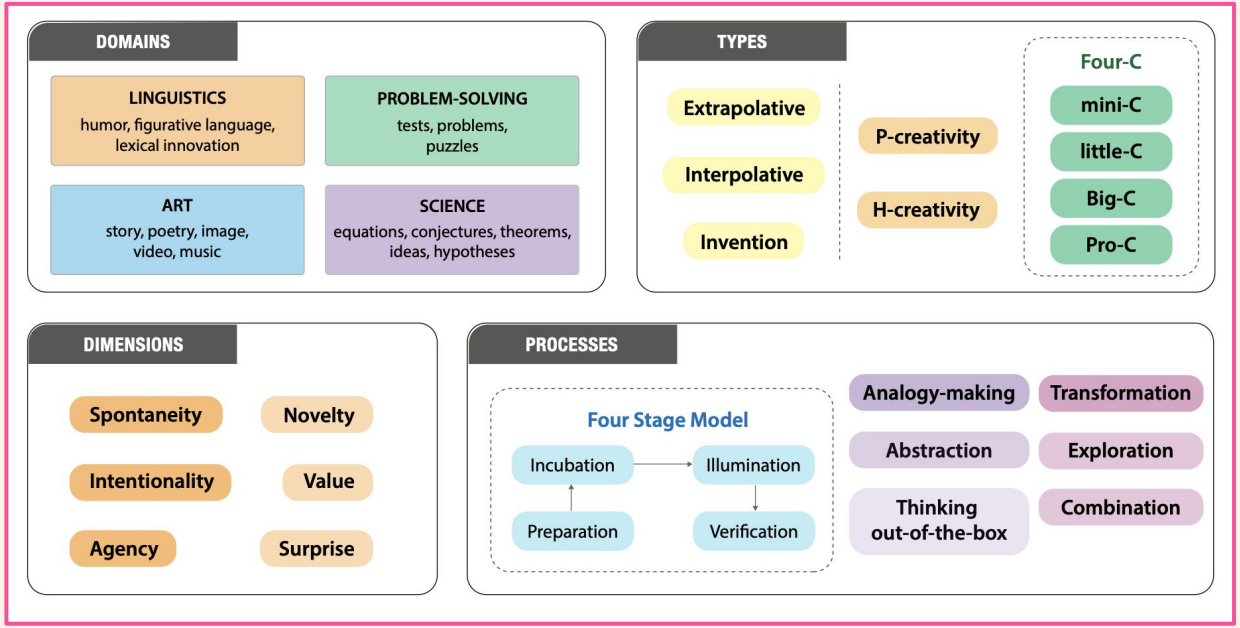
Laboratory Tools

Exp. Chemistry

Equation Discovery

Theorem Proving

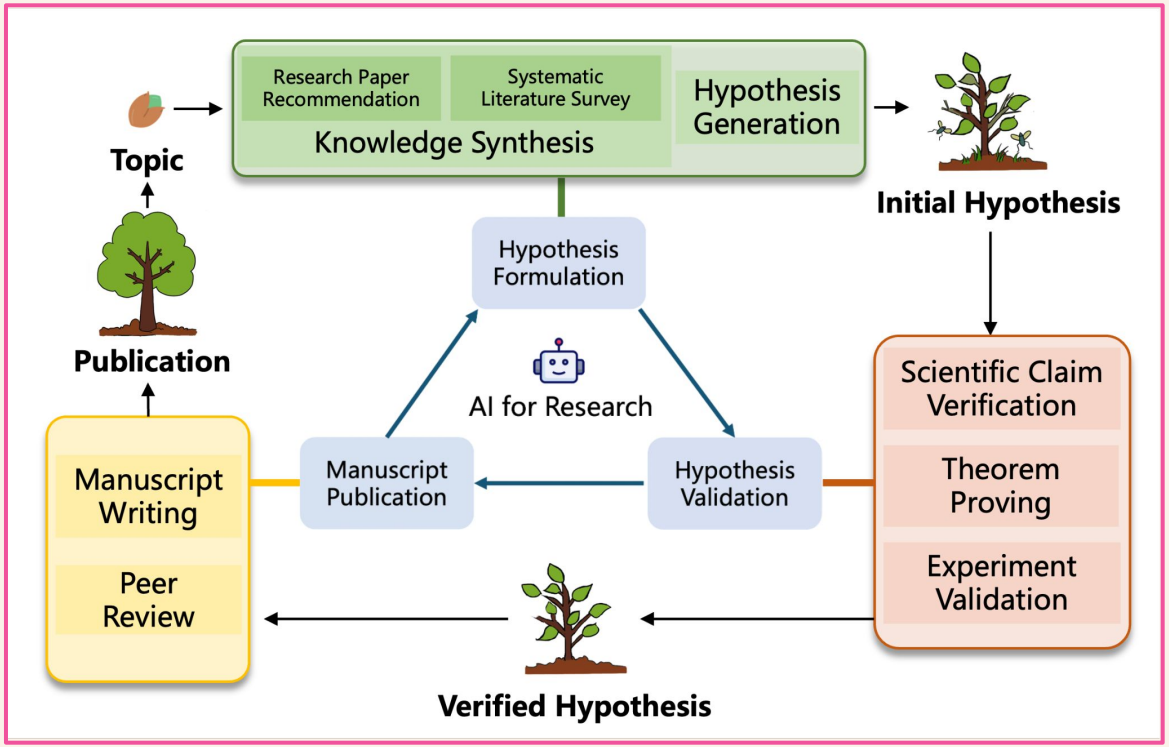
Observational Data*



What are the steps in any Research Process?

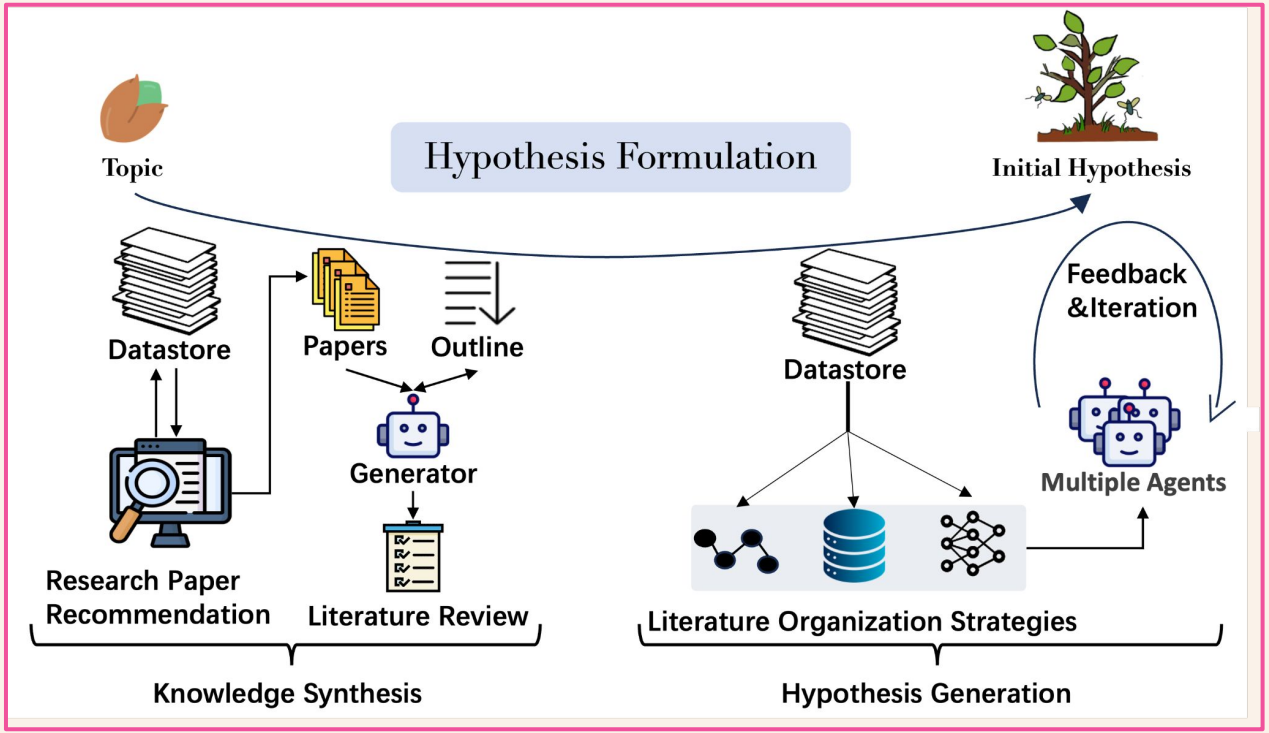
What are different methods of verifying a Hypothesis?

Topic → Hypothesis → Paper



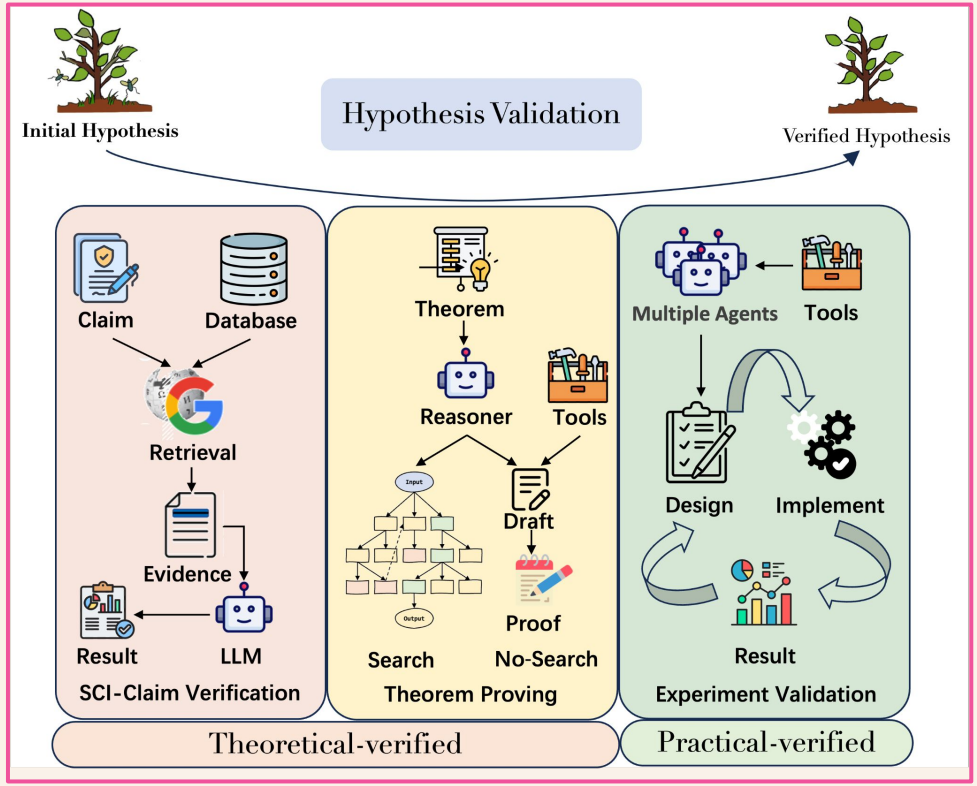
From Hypothesis to Publication: A Comprehensive Survey of AI-Driven Research Support Systems, Arxiv 2025

Hypothesis Formulation



From Hypothesis to Publication: A Comprehensive Survey of AI-Driven Research Support Systems, Arxiv 2025

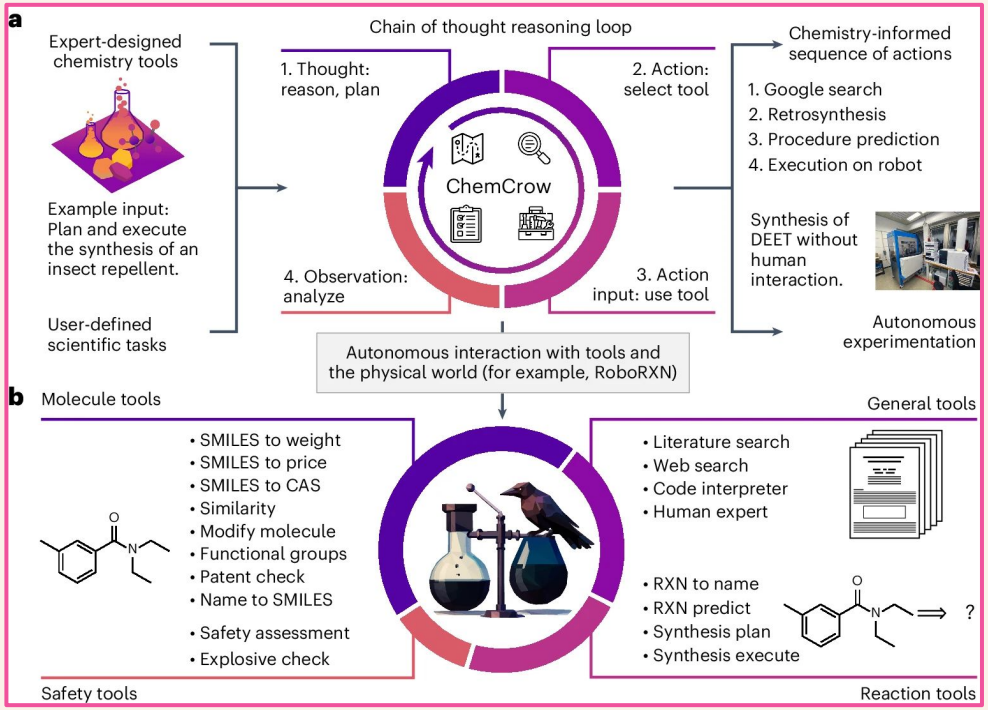
Hypothesis Validation



From Hypothesis to Publication: A Comprehensive Survey of AI-Driven Research Support Systems, Arxiv 2025

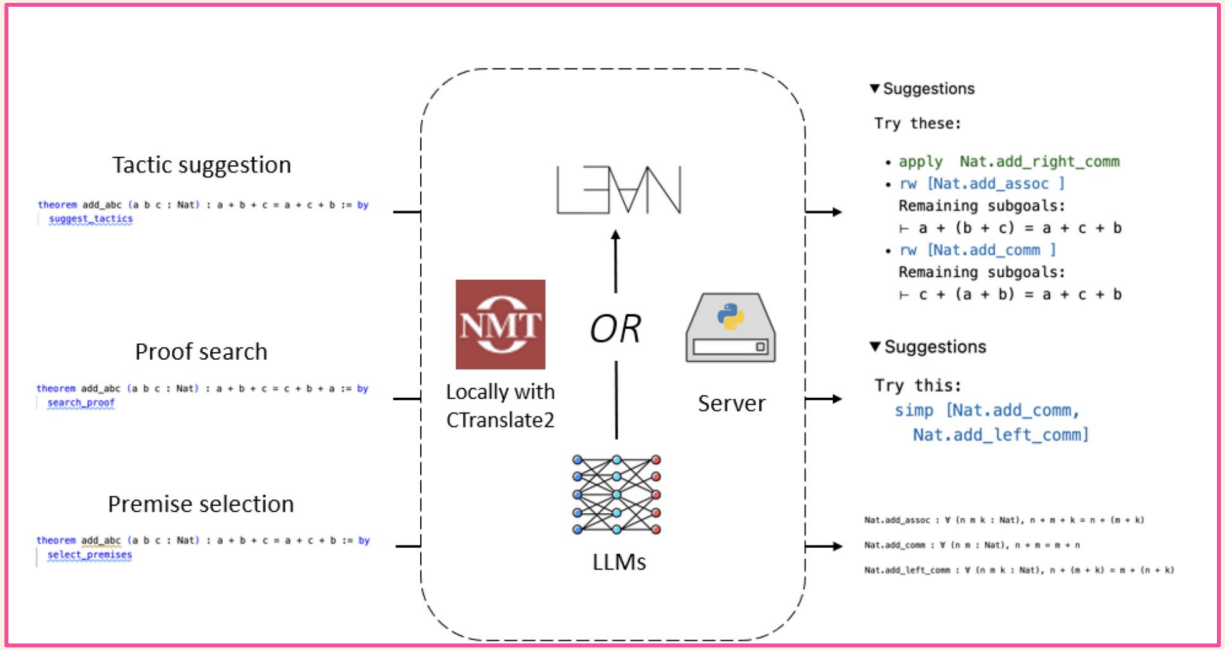
Ai2

LLMs, Lab Tools & Beyond



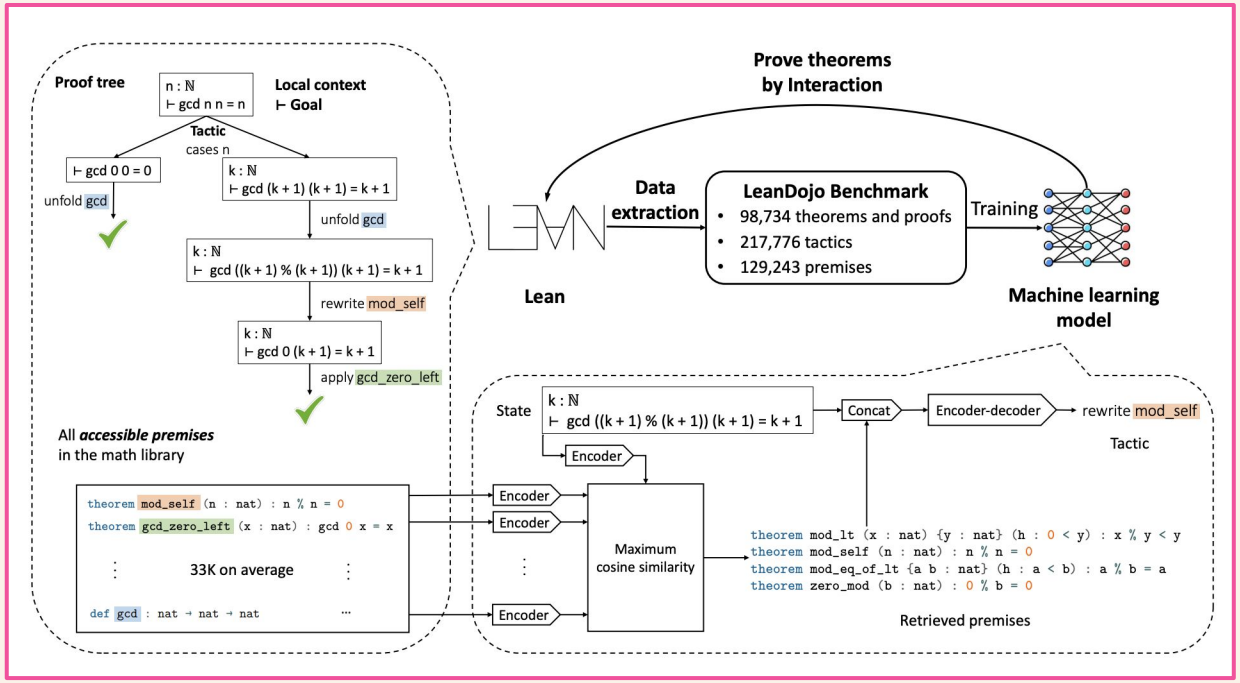
ChemCrow: Augmenting large language models with chemistry tools, Nature 2024

Theorem Proving Copilot



Towards Large Language Models as Copilots for Theorem Proving in Lean, NuerIPS 2023

Theorem Proving Env with LLMs + Retrieval



LeanDojo: Theorem Proving with Retrieval-Augmented Language Models, NuerIPS 2023

Ai2

NVIDIA

MIT Caltech

Methods of Science

Methods of Scientific Inquiry

Theoretical Science

Develop models or theories to explain phenomena

Experimental Science

Conduct experiments to test pre-defined hypotheses

Observational Science

Observe & collect data, build methods to explain it

Methods of Scientific Inquiry

Theoretical Science

Develop models or theories to explain phenomena

Experimental Science

Conduct experiments to test pre-defined hypotheses

Observational Science

Observe & collect data, build methods to explain it

A lot of important science has come out of looking at **observational data.**

National Longitudinal Survey of Youth | 1979

500,000 results in S2 from 1979

37000+ papers published from 1976

Methods of Scientific Inquiry

Theoretical Science

Develop models or theories to explain phenomena

Experimental Science

Conduct experiments to test pre-defined hypotheses

Observational Science

Observe & collect data, build methods to explain it

A lot of important science has come out of looking at **observational data.**

Can we **autonomously** discover

- insights from datasets to reduce turnaround time?
- undiscovered knowledge without performing additional data collection?

National Longitudinal Survey of Youth | 1979

500,000 results in S2 from 1979

37000+ papers published from 1976

Data Driven Scientific Discovery with LLMs

Ai2

Data-driven Discovery

- Comprehensive data-understanding
- Ex-ante hypothesis search/generation
- Planning & orchestrating research pathways
- Execute & verify candidate hypotheses
- Accommodating human feedback
- Reproducible and robust results

Data-driven Discovery: Following Newell & Simon (1976), we define a **heuristic search problem** that aims to describe a given set of observations by uncovering the laws that govern its data-generating process.

E.g., "under context c, variables v have relationship r"

Newell, A. and Simon, H. A. Computer science as empirical inquiry: symbols and search. Commun. ACM, 1976

Data-driven Discovery as a Predictive Task

Given a dataset **D** and a Discovery Goal **G**, derive the most specific hypothesis **H** addressing G and supported by D.

Alternatively,
A data-driven hypothesis **H** is a declarative sentence about the state of the world whose truth value may be inferred from a given dataset D using a verification procedure V: H → {supported, unsupported}, for instance, via *statistical modeling*.

Inspired by Thompson and Skau (2023), we introduce a structured formalism that breaks a hypothesis down into **three hypothesis dimensions**:

Context: **Boundary conditions** that limit the scope of a hypothesis. E.g., "for men over the age of 30"

Variables: **Known set of concepts** that interact in a meaningful way under a given context to produce the hypothesis. E.g., gender, age, or income

Relationship: **Interactions between a given set of variables** under a given context that produces the hypothesis. E.g., "quadratic relationship", "inversely proportional", or piecewise conditionals

W. H. Thompson and S. Skau. On the scope of scientific hypotheses. Royal Society Open Science, 2023

Dataset:

habitat type	nonnative gardening	nonnative unintentional	nonnative agriforest	elevation	...
croplands	5	0	2	675	
wetlands	0	4	1	88	
urban	2	1	0	329	
...	...	...	...	...	

Goal: How did urban land use affect the invasion of different types of introduced plants in Catalonia?

	gold	predicted	score
context	urban habitat type	urban habitat type	1.0
variable	gardening, unintentional	gardening, agriforst	0.3
relationship	reduced	increased	0.0

Final Score: 0.21

Urban land use reduced invasion by gardening plants over unintentionally introduced ones.

Ai2

DiscoveryBench

264 Tasks, 20+ papers, 6 domains

We replicate the **scientific process** undertaken by researchers to search for and validate a hypothesis from datasets

Data-first: Filter papers + workflows based on public datasets: National Longitudinal Surveys, Global Biodiversity Info Facility, World Bank Open Data; 2) replicate in Python.

Replication took up to 90 person-hours per dataset, often (30%) not resulting in success.

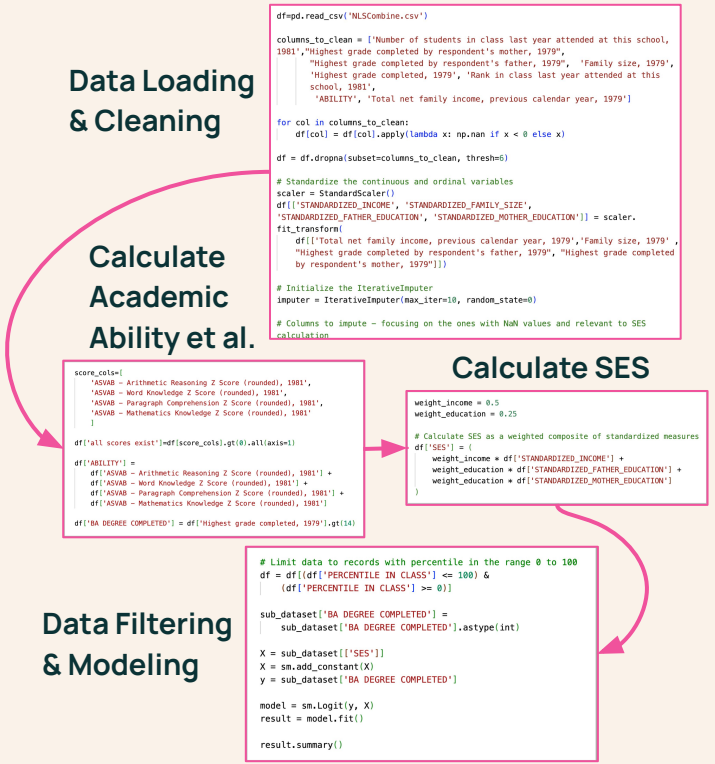
Code-first: Checked 785 repos + datasets, 85% had missing or non-adaptable code to Python, or closed datasets. Only few passed the check.

Papers from Nature, AER, etc.

Task Dataset: Dataset contains information from National Longitudinal Survey of Youth (NLSY79). It includes information about the Demographics, Family Background, Education …

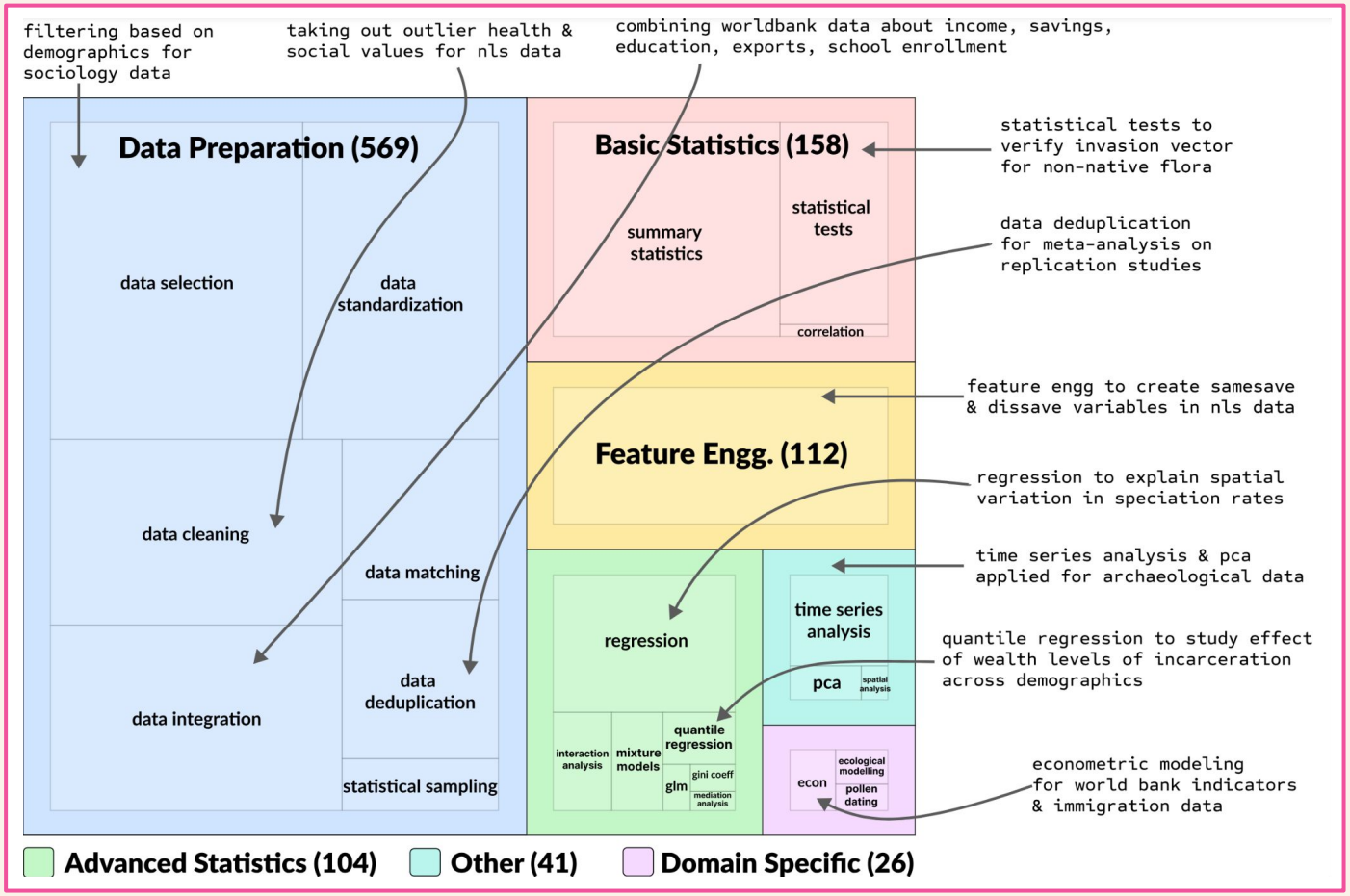
Discovery Goal: How does socioeconomic status affect the likelihood of completing a BA degree?

Target Hypothesis: Socioeconomic status has a positive relationship with college degree completion with a coefficient of 0.4729 with statistical significance.



Ai2

DB-Real (6 domains: sociology, biology, humanities, economics, engineering, & meta-science)



Ai2

Discovery Agents

All discovery agents have access to a python environment, capable of generating and executing programs on the datasets

CodeGen

generates the entire code at one go to solve the task, with help of a demonstration example in the context.

After code execution and based on the result, it generates the NL hypothesis and summarizes the workflow

ReAct

solves the task by generating thought and subsequent codes in a multi-turn fashion.

A traditional sequential-decision maker.

DataVoyager*

is a multi-component data-driven discovery agent.

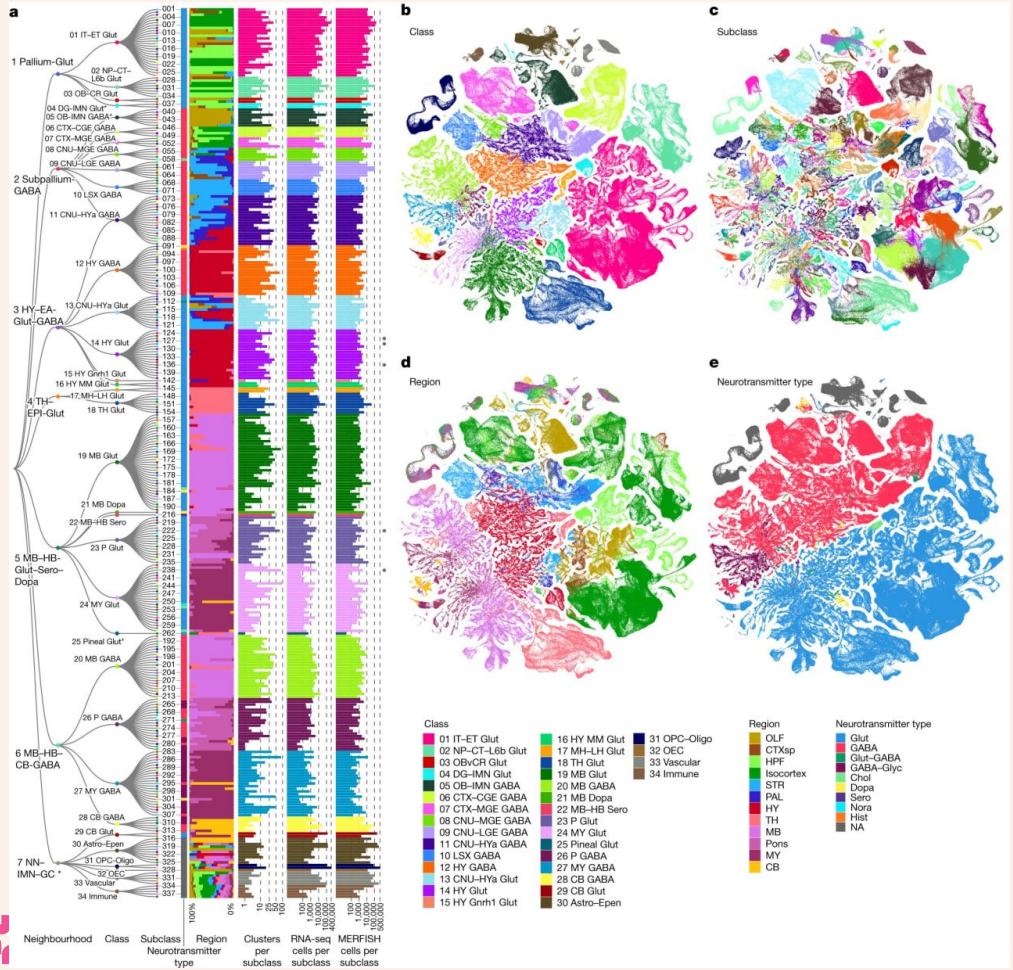
It has four components: planner, code generator, data analysis, and critic, that orchestrate the discovery process.

Reflexion (Oracle)

is an extension of CodeGen agent, where at the end of one trial, we provide an “oracle” feedback about task completion, and it generates a reflection to improve in the next trial till it solves the task, or maximum trials (3) are reached.

A NIMHANS (neuro) Example?

Culmination of decades of work summarized in an image



Landmark Mouse Brain Paper (Ai1)

DataVoyager Prompt:

1. Find the unique classes present in the data.

2. Select the "20 MB GABA" class and plot the distribution of subclasses and neurotransmitters.

3. Select the "23 P Glut" class and plot the distribution of subclasses and neurotransmitters.

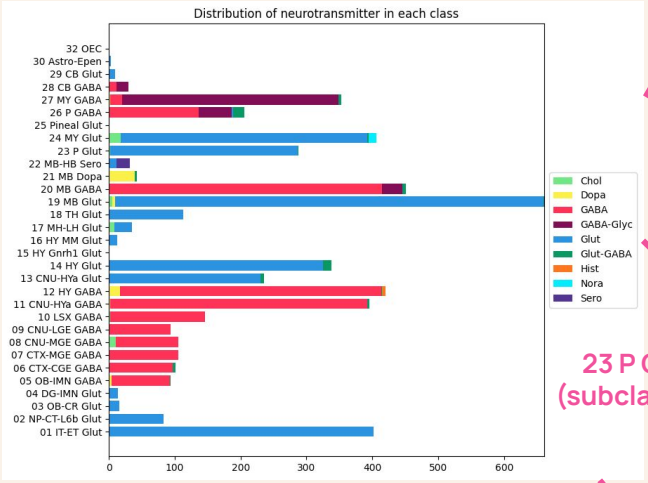
4. Plot the supertypes and neurotransmitters distribution for the same filtered class.

5. Interpret the unique supertypes for this filtered class.

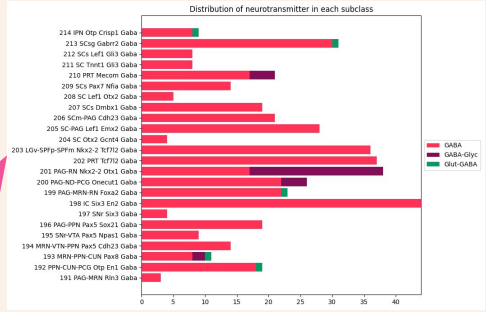
A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain
https://www.nature.com/articles/s41586-023-06812-z

DataVoyager Traces for WMB Data

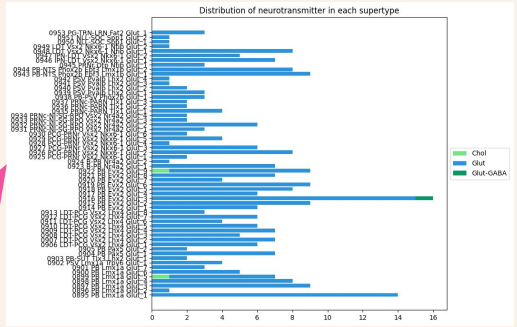
WMB Classes by nt



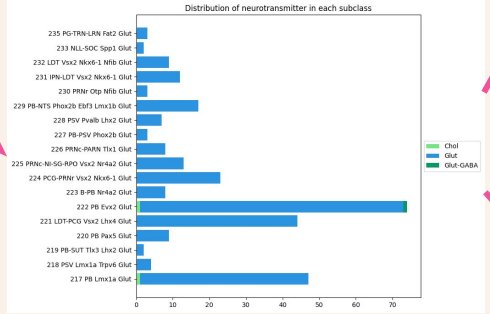
20 MB GABA (subclasses)



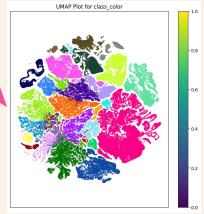
Supertypes



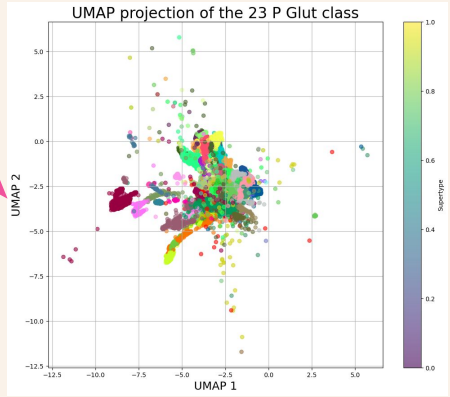
23 P Glut (subclasses)

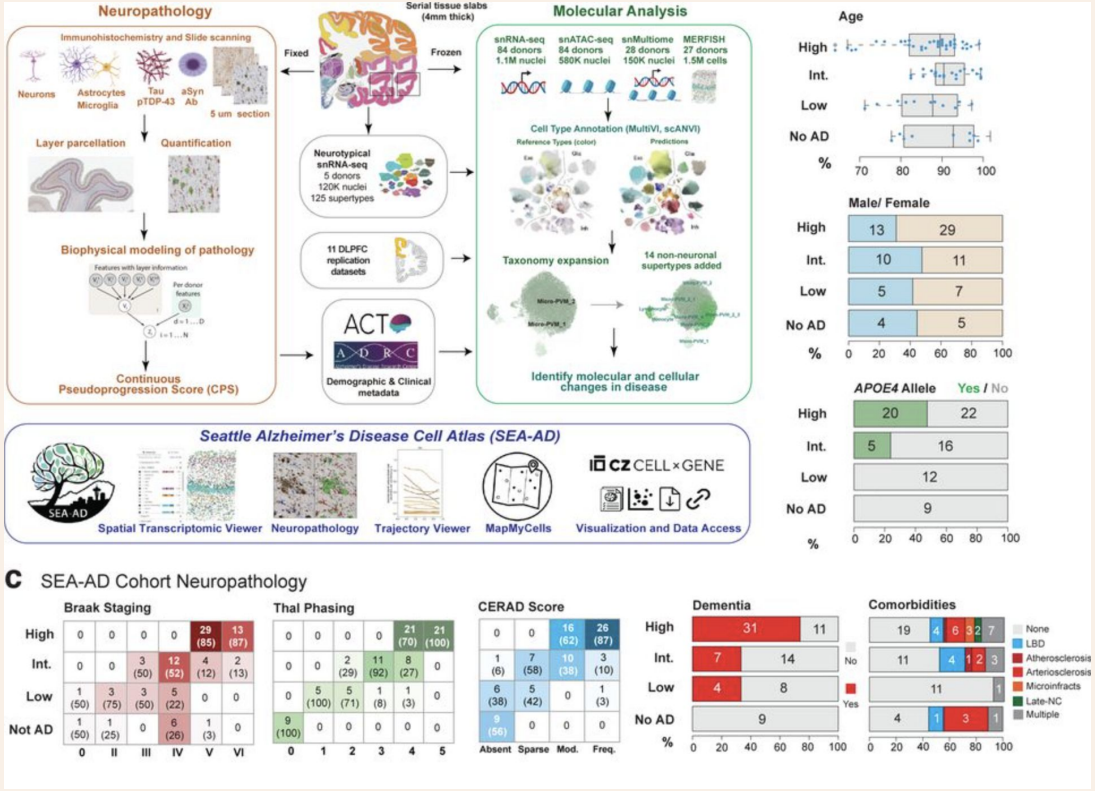


Projection



Projection





Integrated multimodal cell atlas of Alzheimer's disease

Data Input: Not sure of exact data used in paper or preprocessing. We currently used the following:

Donor Data

66 columns, and info includes metadata for each donor such as 'Primary Study Name', 'Age at Death', 'Sex', 'Race', 'CERAD score', 'Overall CAA Score', 'Highest Lewy Body Disease', 'Total Microinfarcts', 'Atherosclerosis', 'Arteriolosclerosis', 'LATE', 'RIN', and whether the donor is 'Severely Affected'.

MTG Data

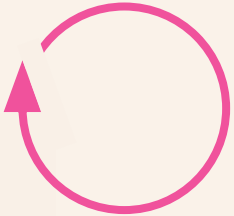
394 columns, measurements related to AT8 positive areas in different layers of Grey matter, pTDP43 positive areas, and other neuropathological quantifications for each donor.

DataVoyager is a part of **Asta** - a Science Copilot

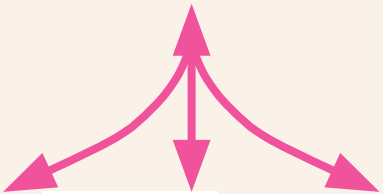
1) Collaborates naturally with human scientists

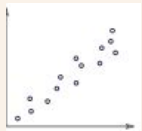
2) Uses tools on humans' behalf

3) Learns & improves over time.



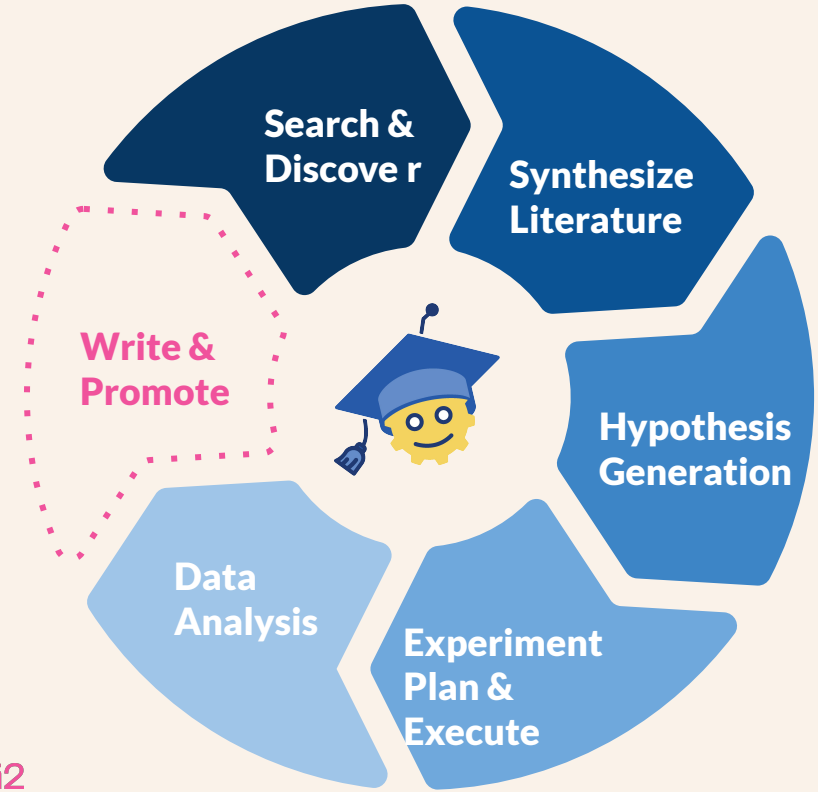
Slack
Semantic Reader
Jupyter





… with Infra. for Personalization & Optimization

Research Functions



Surfaces

Slack
Web
Semantic Reader
Jupyter

Interaction Store

Evaluation Harness

Some Accolades

New Directions in the Area Influenced by our Work

BLADE: Benchmarking Language Model Agents for Data-Driven Science

POPPER: Automated Hypothesis Validation with Agentic Sequential Falsifications

ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery

QR Data: Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data

Recognized with other major works including the aforementioned Nobel Prize in 2024

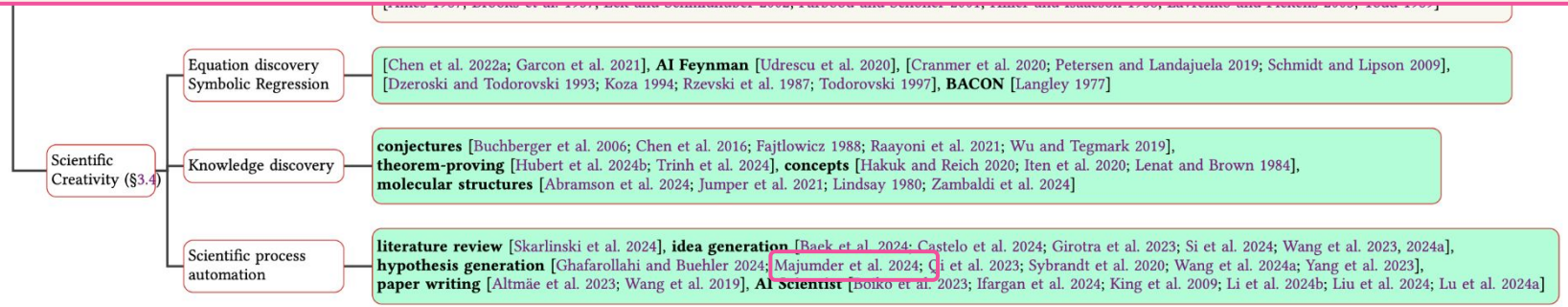


Fig. 2. Taxonomy of creativity in AI covering areas of linguistic creativity, creative problem-solving, artistic and scientific creativity. Note that this taxonomy is not exhaustive, but rather a representative view of the key works.

***Creativity in AI: Progresses and Challenges**, EPFL Zurich. Mete Ismayilzada, Debjit Paul, Antoine Bosselut, Lonneke van der Plas*

Learnings

Learnings for in LLMs + Open Source

Instrument the **Data Pipeline**, Not Just the Model

Building **Evaluations** that Actually Matter

Adopting & **Extending Frameworks** (LangChain, AutoGen, etc.)

Treat **Prompts** and Pipelines as **Code Assets**

Exploit Small Models First, Big Models Last

Exploit Big Models First, Small Models Last

A/B (or **Ablations**) Everything Behind a Feature Flag/ Config

How to contribute to fundamental AI from India?

What happens with petabytes of data?

Or millions of lines of code?

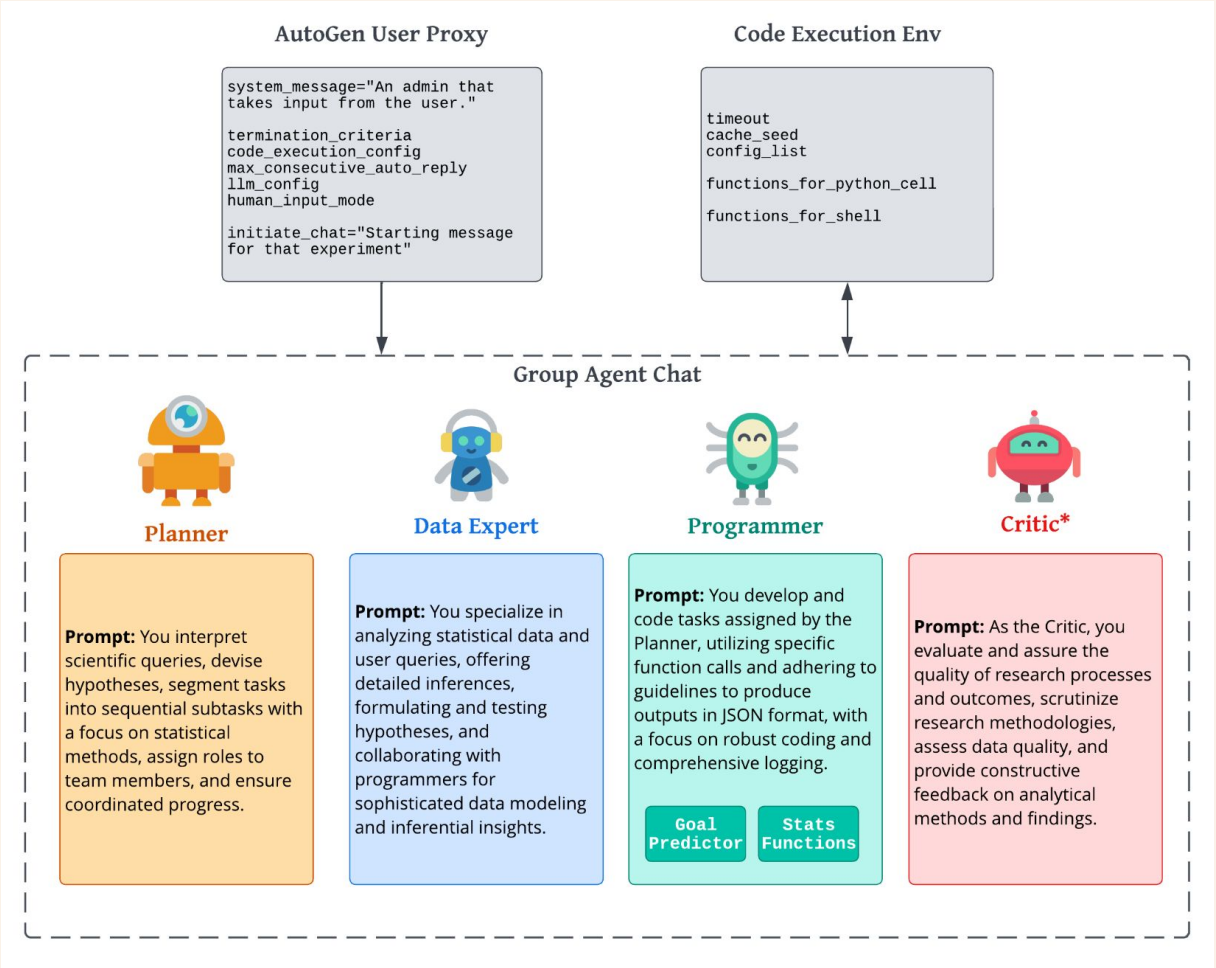
Appendix

DataVoyager

All of the subagents and the controller is based on with a specialized prompt.

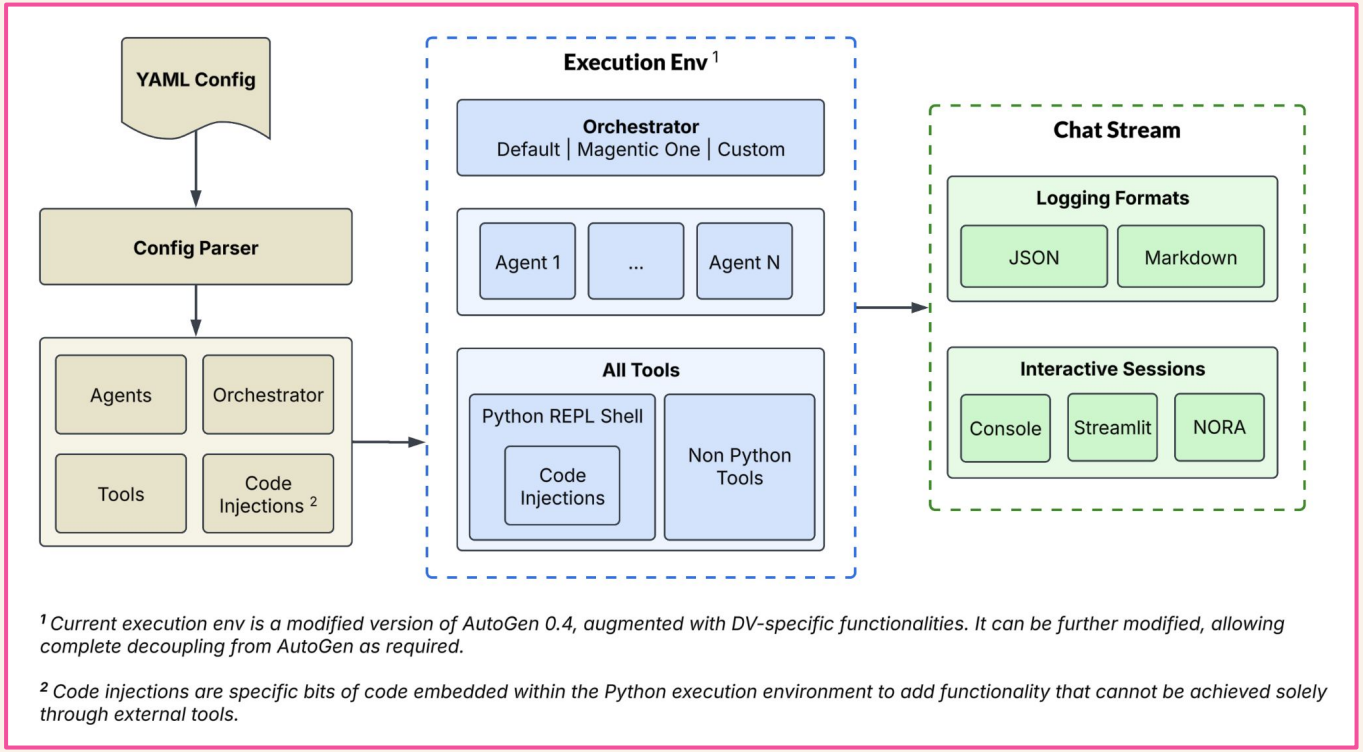
They see the full history, but the agent-specific prompt elicit specialized behavior.

Programmer has the function-calling ability which allows it to “execute” the generated code in an **interactive python shell**.



Ai2

DataVoyager Core Architecture



[1] *Current execution env is a modified version of AutoGen 0.4, augmented with DV-specific functionalities. It can be further modified, allowing complete decoupling from AutoGen as required.*

[2] *Code injections are specific bits of code embedded within the Python execution environment to add functionality that cannot be achieved solely through external tools.*

LLM Benchmarks

- DataGLUE
- NORA - End to End Science
- Bio - Neuro/ Immunology

Better Agents/ Models for

- Data Driven Discovery
- End to End Science
- Neuro/ Immunology/ Genetics

Collabs

Auto-exploration with UMass/ Andrew McCollum group

Better memory for coding/ agents with Zurich U

Better Cost Efficiency

Pareto optimal model usage/ model routing for DDD

100X+ difference b/w cheaper & advanced models

Related Projects List

Augmenting code editing with AI features

- Fork open-source AI code editors like melty, zed or pearai or aider
- Add new features like AI diagraming for architecture etc. in them
- https://aider.chat/docs/leaderboards/

Enterprise heavy requirements

- Add guardrails to prevent certain data leakage
- Serve data while making sure it is not hallucinated
- Can be RAG or database query driven

Other Ideas

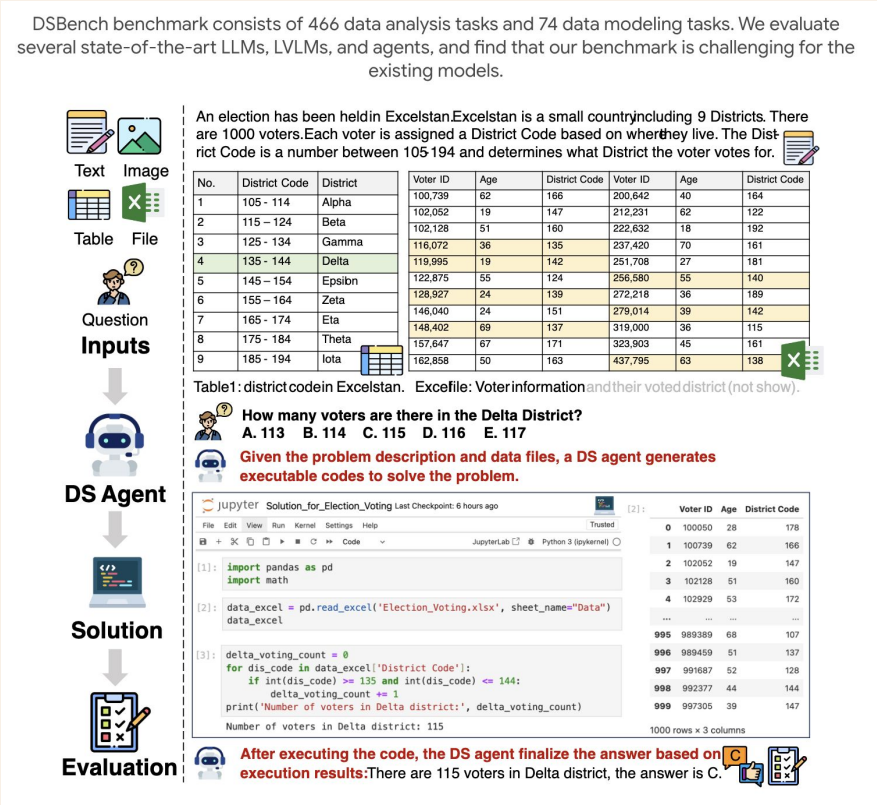
- AutoRAG
- RAG + agents on domain specific areas like medicine
- RAG + agents for indian docs
  - Analyze data
  - Make it available in Indian languages

More Benchmarks

DSBench

https://github.com/LiqiangJing/DSBench

- Has files such as .txt, .xlxs
- Has metadata/background about the data
- MCQ
- Requires computation
- Autogen-based GPT-4 achieves 87.84% task success

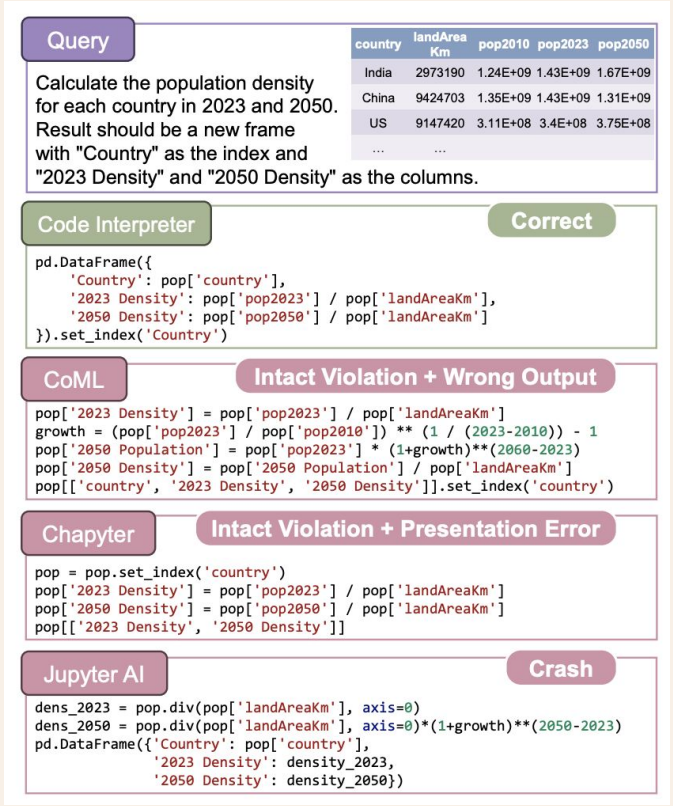


More Benchmarks

DSEval

https://github.com/MetaCopilot/dseval

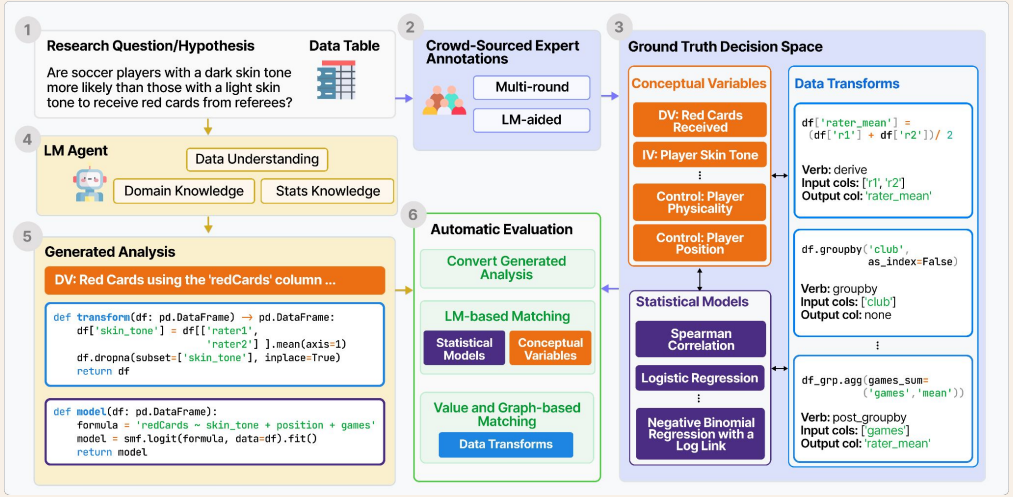
- Has files such as .txt, .xlxs
- Standard datasets like titanic, twitter
- QA with no options
- Requires computation
- GPT-4-based agents reach ~70% pass rate
- Has other evaluations/validators



More Benchmarks

BLADE
https://github.com/behavioral-data/BLADE/tree/main

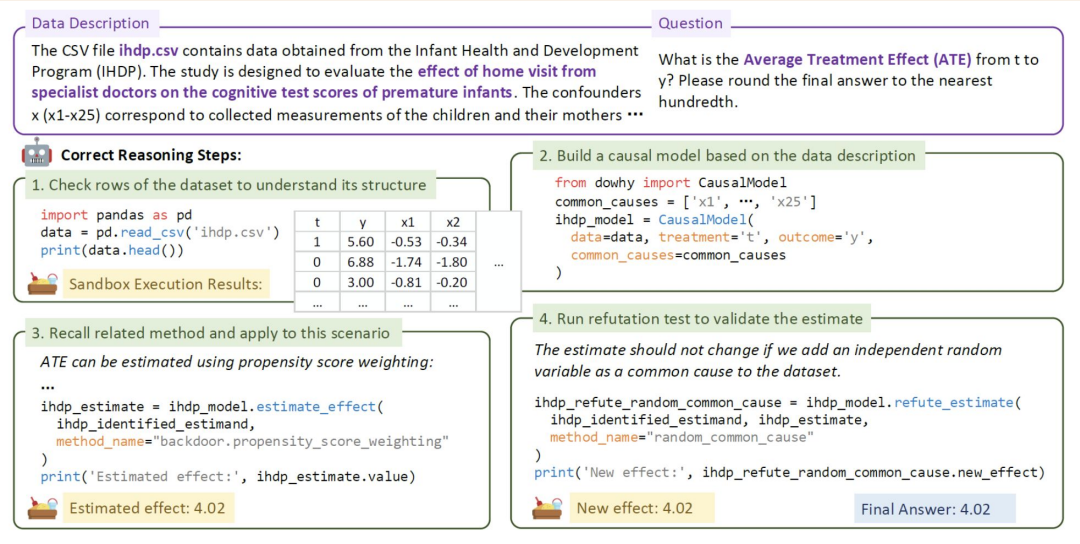
- Has files, .csvs
- Has metadata about the datasets
- Research Q/Hypothesis as a goal (most in the format: Is this true?)
- MCQs
- Requires computation



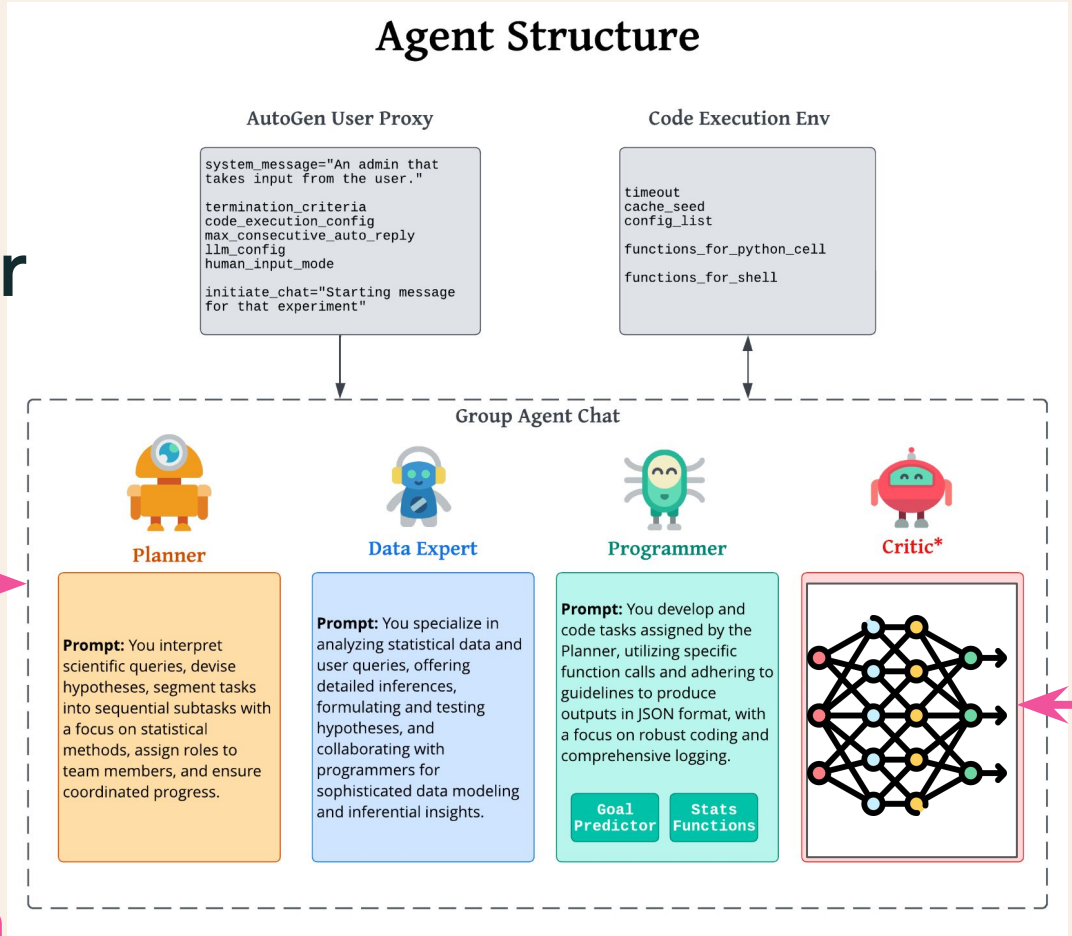
More Benchmarks

QRData
https://xxxiaol.github.io/QRData/

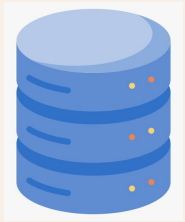
- Subset of benchmark has files, .csvs
- MCQ
- Requires computation
- GPT-4 achieves ~60% accuracy



Memory with DataVoyager



Project Memory



Library functions,
User preferences,
Domain knowledge,
Explored hypotheses

Private to project
(can be proprietary info)

Finetuned critic

Learned from a big corpus/ Voluntarily shared logs

But, what if the goal is unknown?

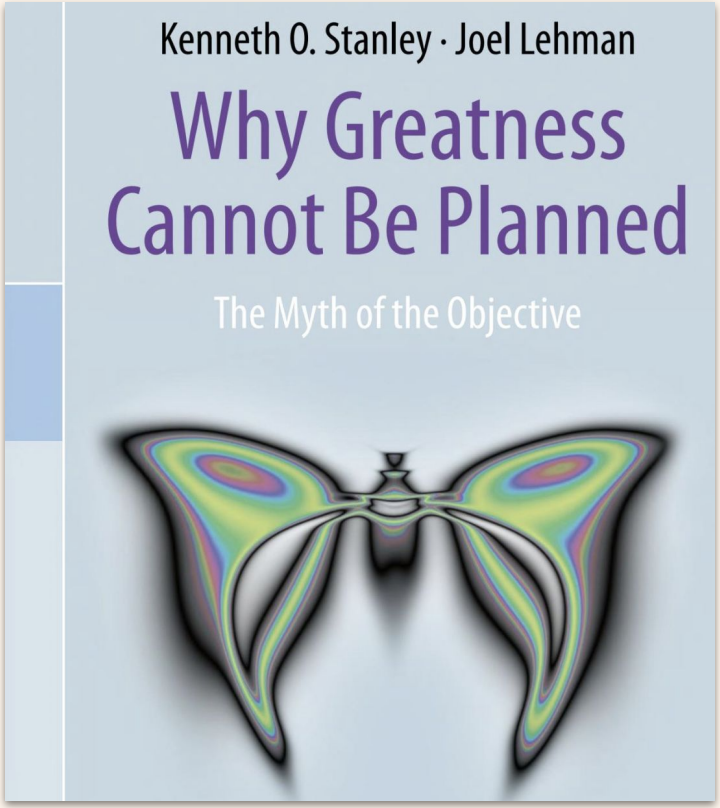
Can **open-ended exploration** lead to **useful discoveries**?

- No explicit research **question** or discovery **goal**
- Only given a dataset
- Optionally, a set of known hypotheses

- Should align with human notions of **"interestingness"**
  - Discoveries made may answer human-posed research queries in the future
- Serves as a **building block for future discoveries**

Ai2

Not just a philosophical point



Could address **practical limitations** in **goal-driven discovery:**

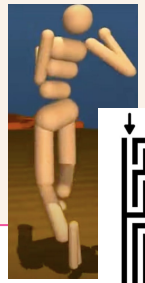
- **Low dataset coverage:** Performs shallow analyses; Struggles to sufficiently cover the dataset; misses promising experiment designs.
- **Low diversity:** Repeated execution does not produce diverse outputs. Inference-time scaling laws do not seem to hold in our discovery setting.

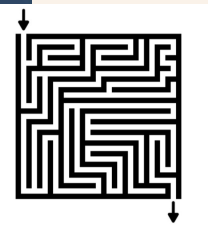
Ai2

Open-endedness in data-driven discovery

Benefits:

- **Immediate feedback:** Closed loop of proposing and executing experiments on a given dataset allows for a continual, autonomous process to collect verifiable hypotheses.
- **Expanding search space:** As hypotheses are collected, new opportunities arise to further explore known hypotheses, combine them, or continue sampling new ones.
    - Solutions quickly plateau in static search spaces. E.g., maze solving, bipedal walking.

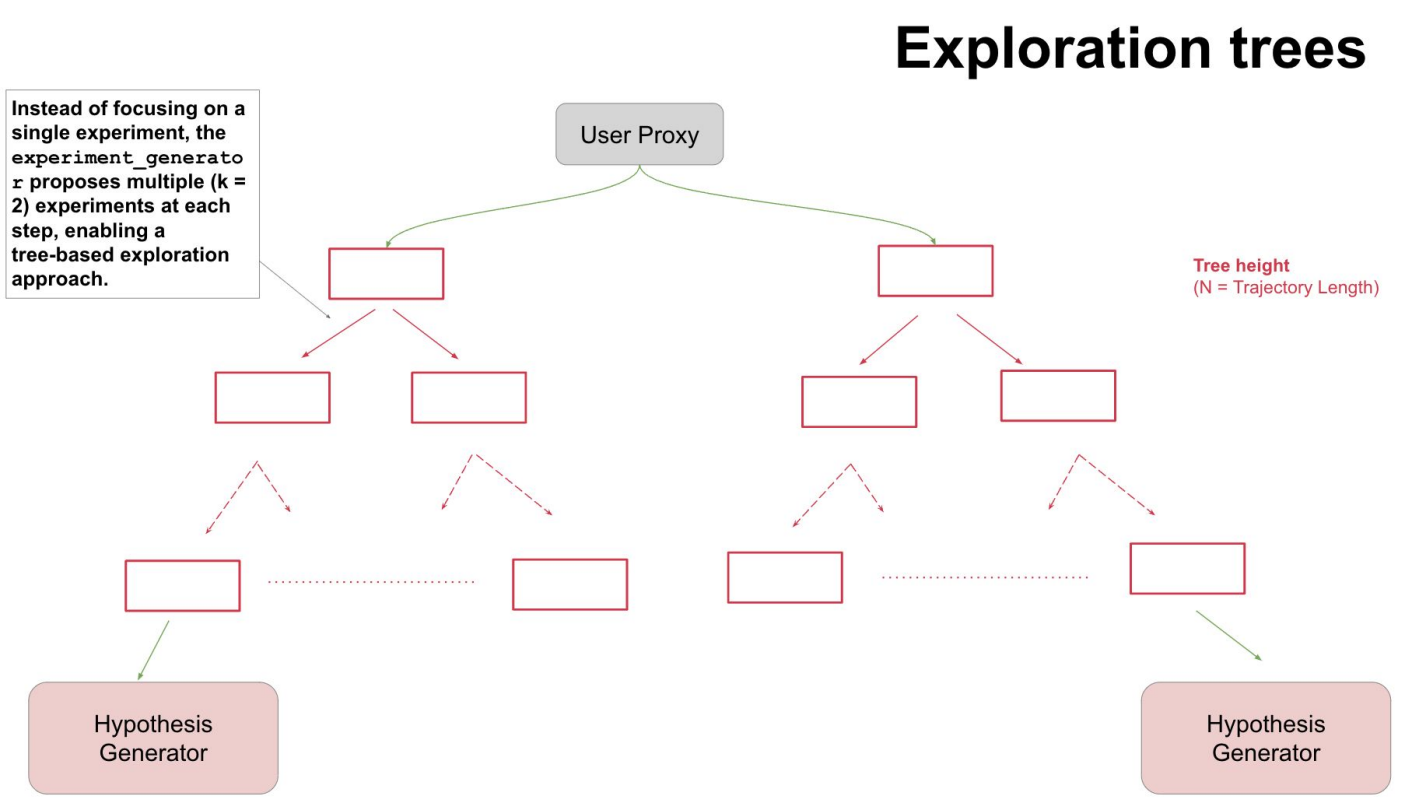




3 main challenges:

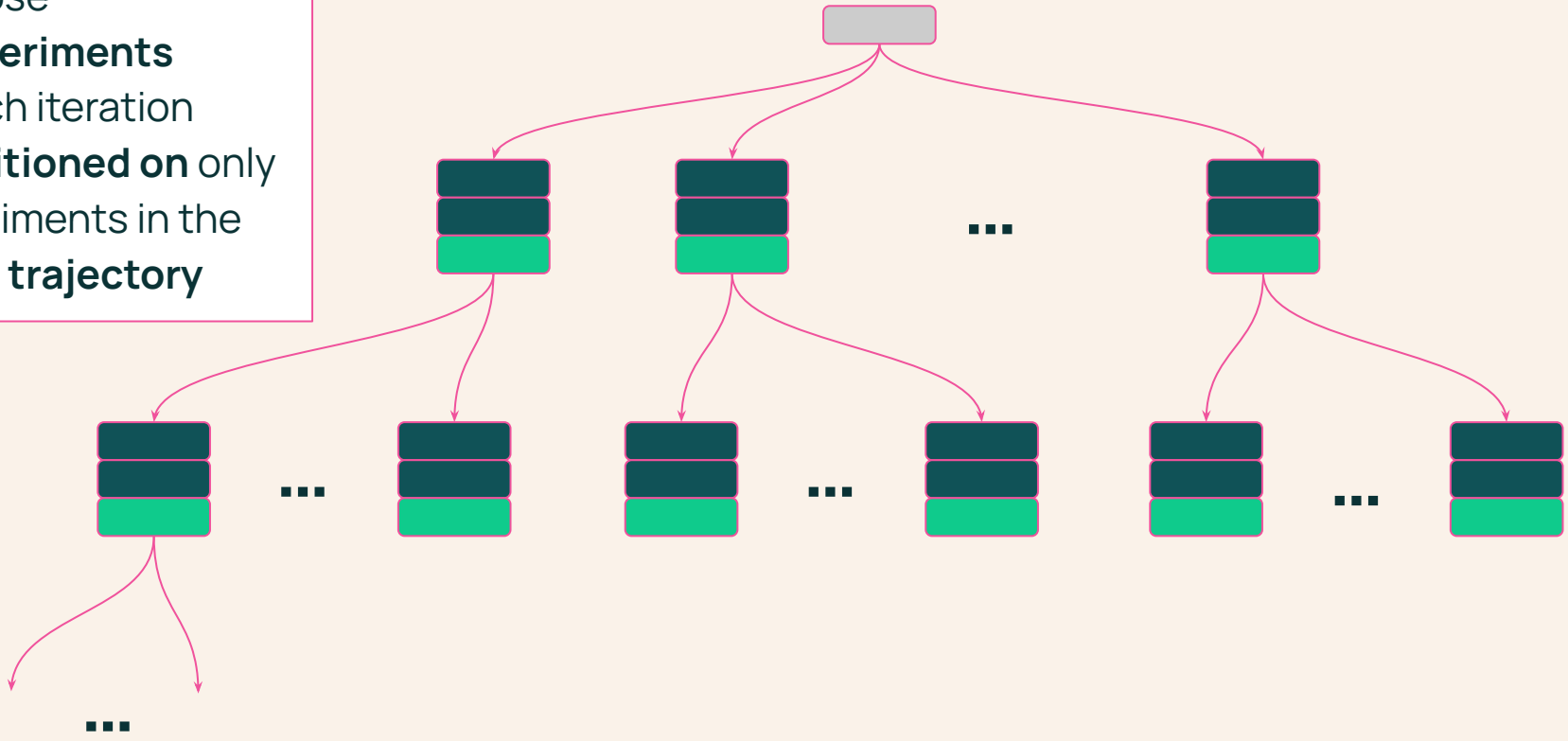
1. Repeated sampling is insufficient. **Need a method to sample diverse trajectories.**
2. Is diversity/novelty a sufficient metric? **What reward function should be used to guide exploration?**
    - Interestingness? Utility? **Hard to define!**
3. Even with useful reward functions, **how should models be steered to continually generate interesting hypotheses**, especially given the dynamical nature of “interestingness”?
    - **Is prompting enough? Can we do more?**

AutoDataVoyager - Exploration Trajectories



AutoDV: Experiment Tree

Propose
k **experiments**
in each iteration
conditioned on only
experiments in the
same trajectory



Ai2

Experiment Tree: Samples

E: Analyze the influence of family size on educational achievement, specifically focusing on the completed BA degree status.
R: Significant difference in family size between those who completed a BA degree and those who did not, with smaller family sizes associated with degree completion.
H: Individuals from smaller families are more likely to complete a BA degree compared to those from larger families, as evidenced by the significant difference in mean family sizes between the two groups.

E: Investigate potential nonlinear relationships between socioeconomic status (SES) and academic performance indicators such as ASVAB scores and class percentile.
R: Slight nonlinear relationship with SES, but linear relationship remains dominant pattern.
H: There is a slight nonlinear relationship between socioeconomic status (SES) and academic performance indicators, with a quadratic term indicating diminishing returns at higher SES levels. However, the linear relationship remains the dominant pattern, suggesting that SES consistently predicts academic performance across its range.

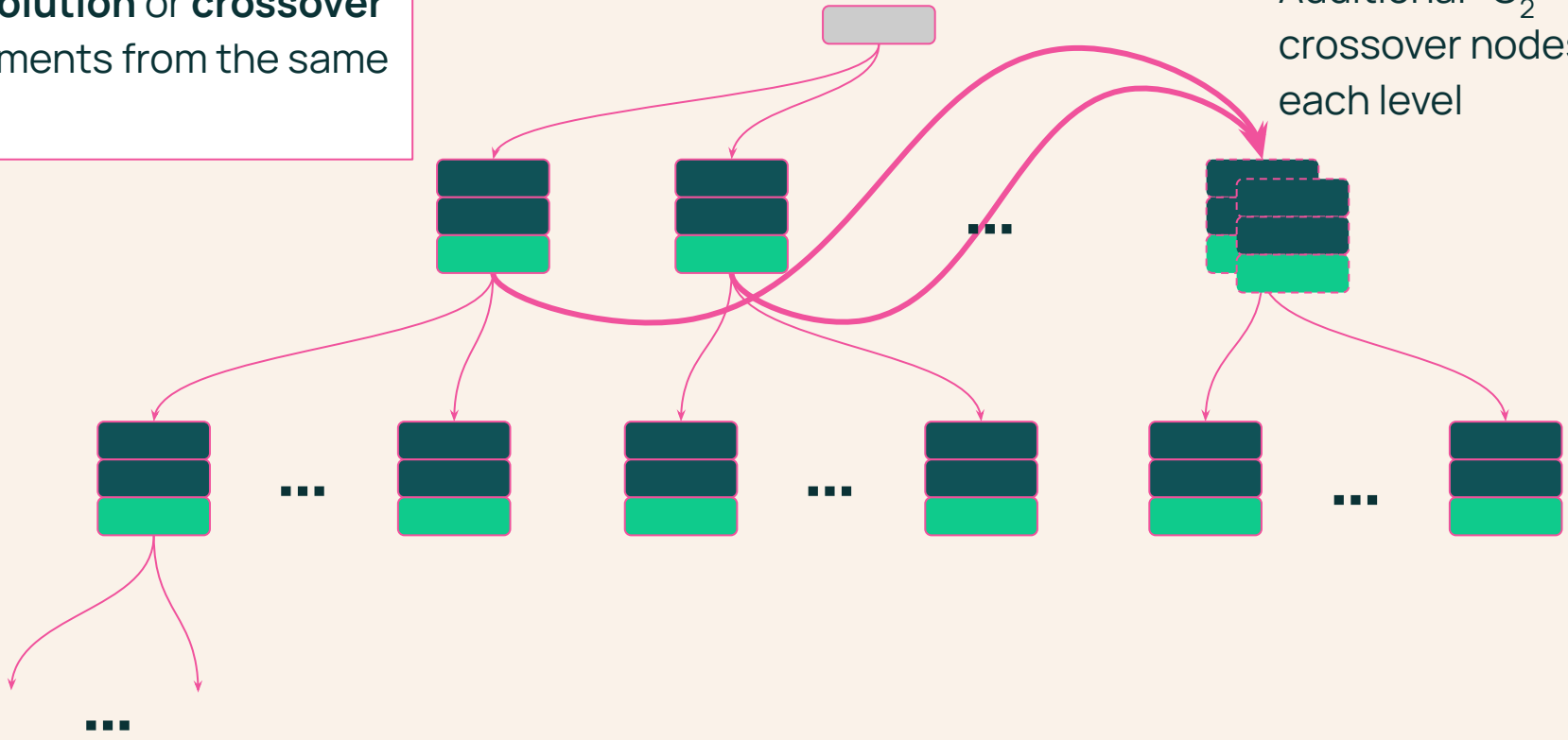
E: Conduct a clustering analysis to identify distinct profiles of students based on their ASVAB scores, class percentile, SES, race, and gender.
R: Three distinct clusters identified, showing varying academic performance and SES levels.
H: Distinct profiles of students can be identified based on ASVAB scores, class percentile, and SES, with higher SES associated with better academic performance. These profiles may also reflect racial and gender disparities, with certain groups more likely to belong to clusters with lower academic performance.

Higher diversity and **higher complexity in experiment design.**

AutoDV: Crossovers

Co-evolution or **crossover** experiments from the same level

Additional ${}^{n}C_{2}$ crossover nodes in each level



Ai2

Walkthrough of the DataVoyager system

https://x.com/surana_h/status/1786097912147239157

https://x.com/mbodhisattwa/status/1761061506127655244

LLMs have revolutionized NLP.

But they often fall short in key areas.

LLM Limitations (for Advanced Tasks)

- Hallucinations
- Limited context
- Loses memory over longer context
- Non-trivial to verify tasks
- Opaque
- Poor adaptability for out of distribution domain & tasks
- Cannot integrate into specific user workflows

LLM Limitations (for Advanced Tasks)

- Hallucinations
- Limited context
- Loses memory over longer context
- Non-trivial to verify tasks
- Opaque
- Poor adaptability for out of distribution domain & tasks
- Cannot integrate into specific user workflows

Cool demo, but how do I *use* it?

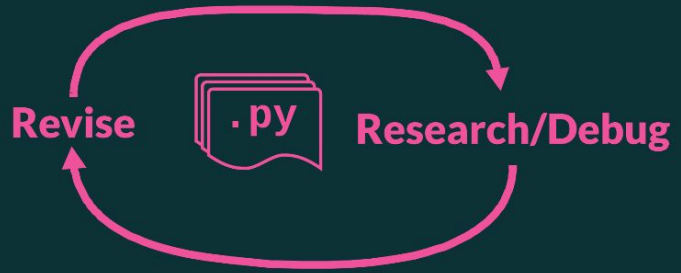
An Example for NLP Code

Non-Agentic Workflow

Write a Python script that uses a Hugging Face model for sentiment analysis on a given bio dataset. Please type out the code in one go without waiting or even using the backspace.

Agentic Workflow

- Begin by refining the problem scope into a detailed specification
- Iterate with domain experts: doctors, clinical data scientists
- Design a modular preprocessing pipeline that can adapt to new domain rules
- Start with a baseline model like BioBERT and iterate
- Define domain-specific metrics that matter including accuracy & F1-score
- Evaluate results on the defined metrics & iterate the code



What are the solutions?

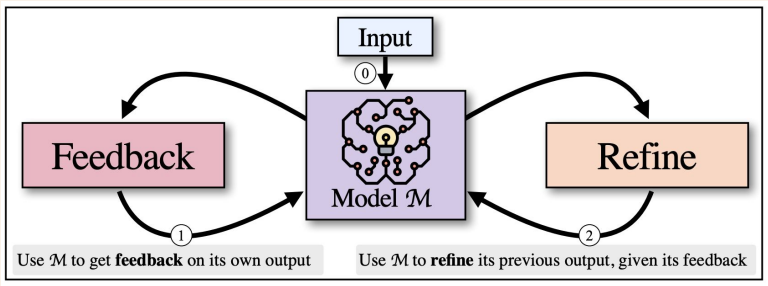
Agentic Design Patterns can tackle the limitations to make more productive use of the LLMs.

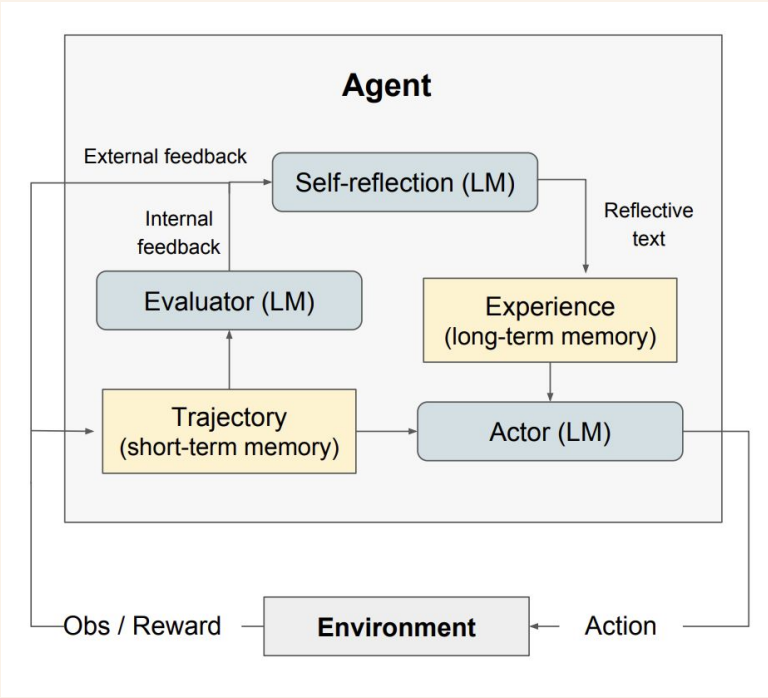
1. Reflection
2. Tool use
3. Planning
4. Multi-agent collab!

Reflection

Verify & reflect the LLM output by external feedback (i.e. unit tests) & LLMs. Use the reflection to iterate the results.

- Self Refine
- Reflexion





Self-Refine: Iterative Refinement with Self-Feedback. *Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, Peter Clark. NIPS 24.*

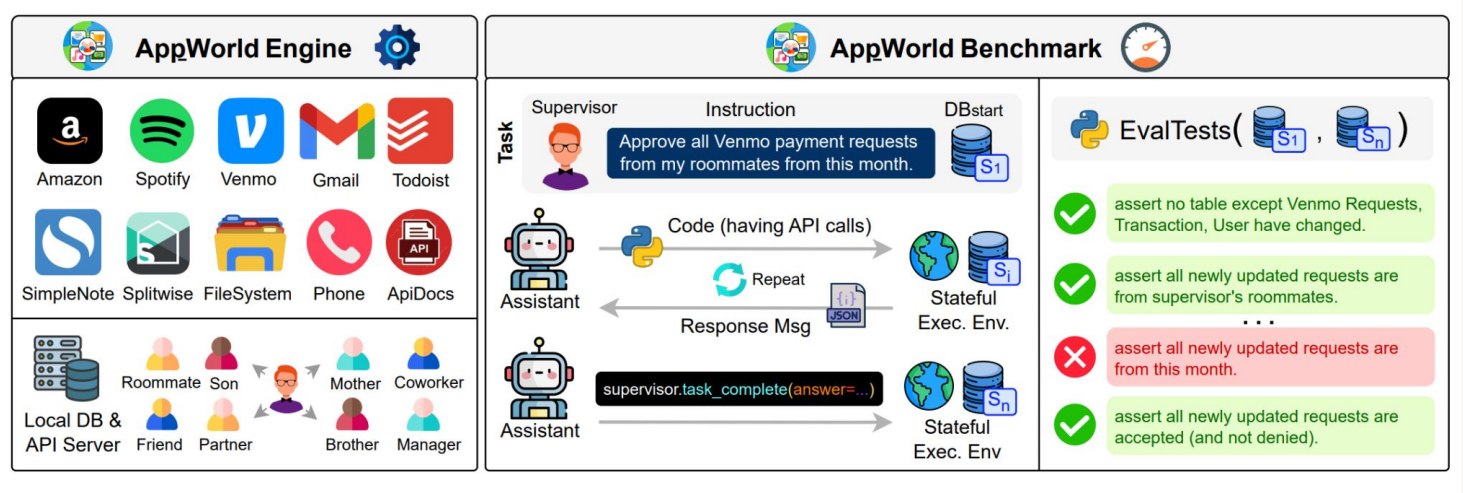
Carnegie Mellon University · University of Washington · Ai2 · Google · NVIDIA

Tool Use

Connect with various tools & functions like:

Browser, APIs, code exec, **custom code** (domain specific code hard for LLMs to generate), search engine etc.

- AppWorld
- Gorilla LLM Exec
- Function calling

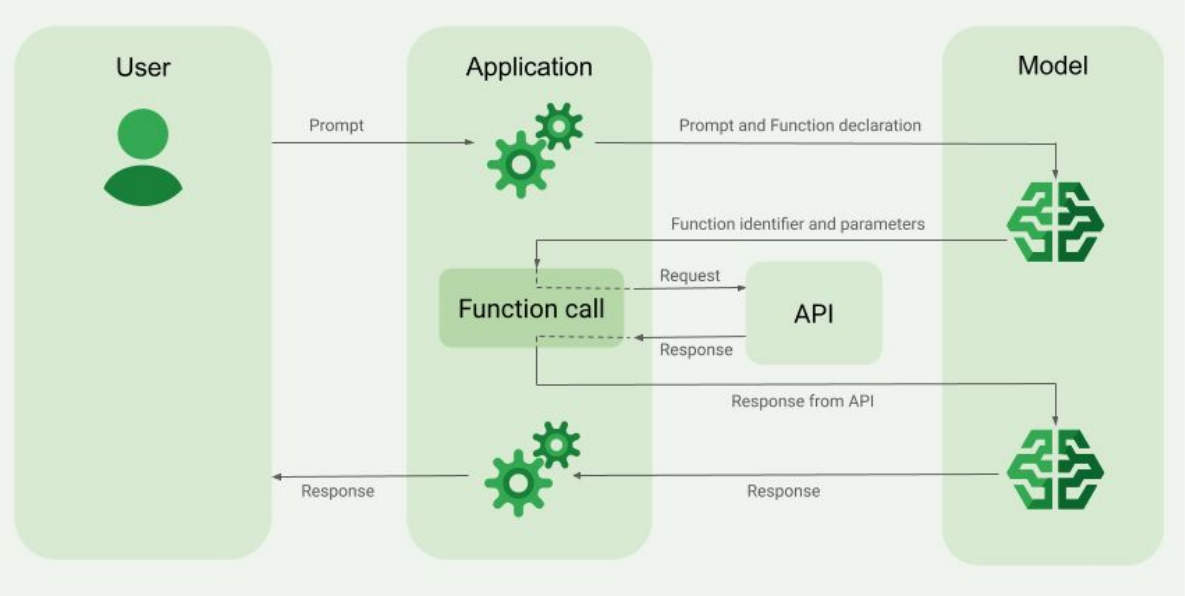


Tool Use

Connect with various tools & functions like:

Browser, APIs, code exec, **custom code** (domain specific code hard for LLMs to generate), search engine etc.

- AppWorld
- Gorilla LLM Exec
- **Function calling**



** showcased toolcalling is from Google Gemini - but a similar function call works across platforms*

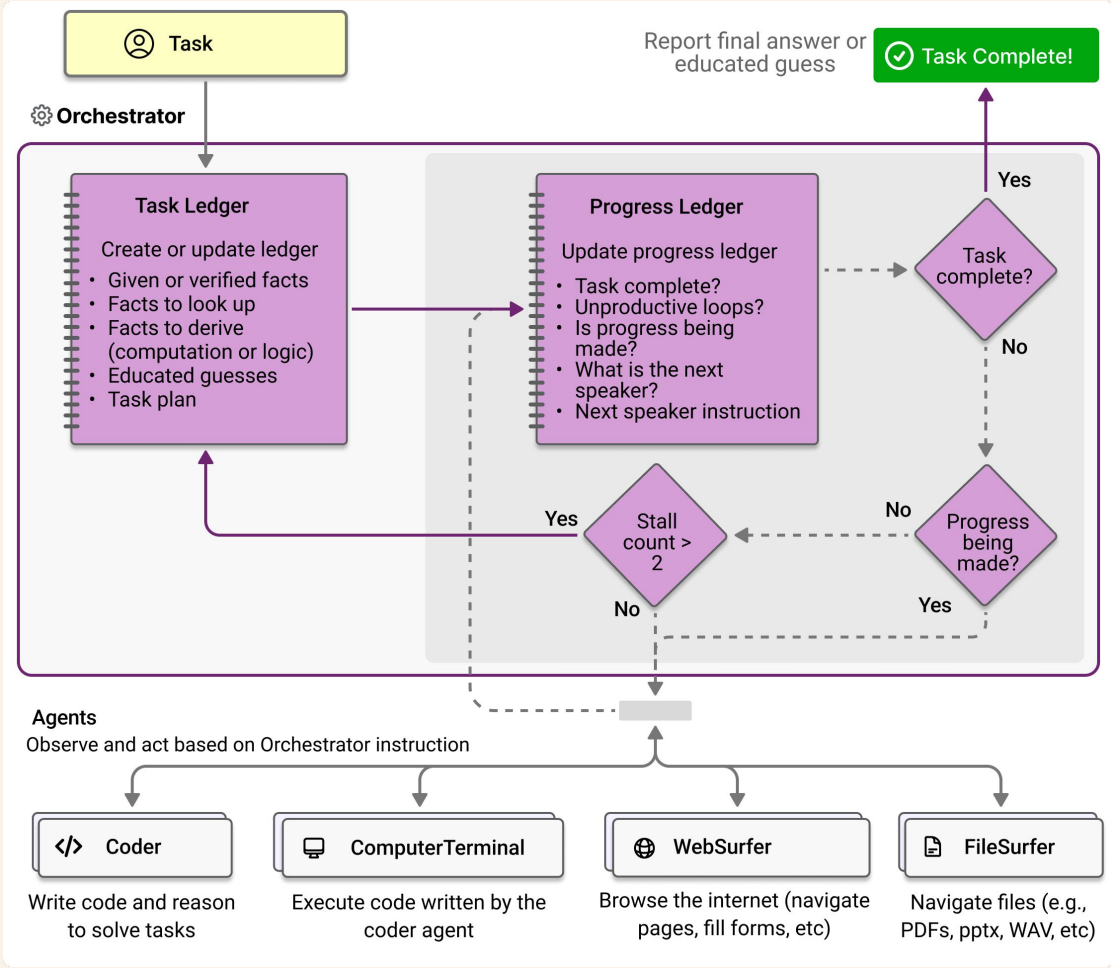
Ai2

Planning

Plan tasks, track them while performing a task and reason over them if needed.

- Autogen Magento
- MetaGPT
- OpenHands
- HuggingGPT

Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. *Fourney et. al Microsoft Tech Report 2024.*

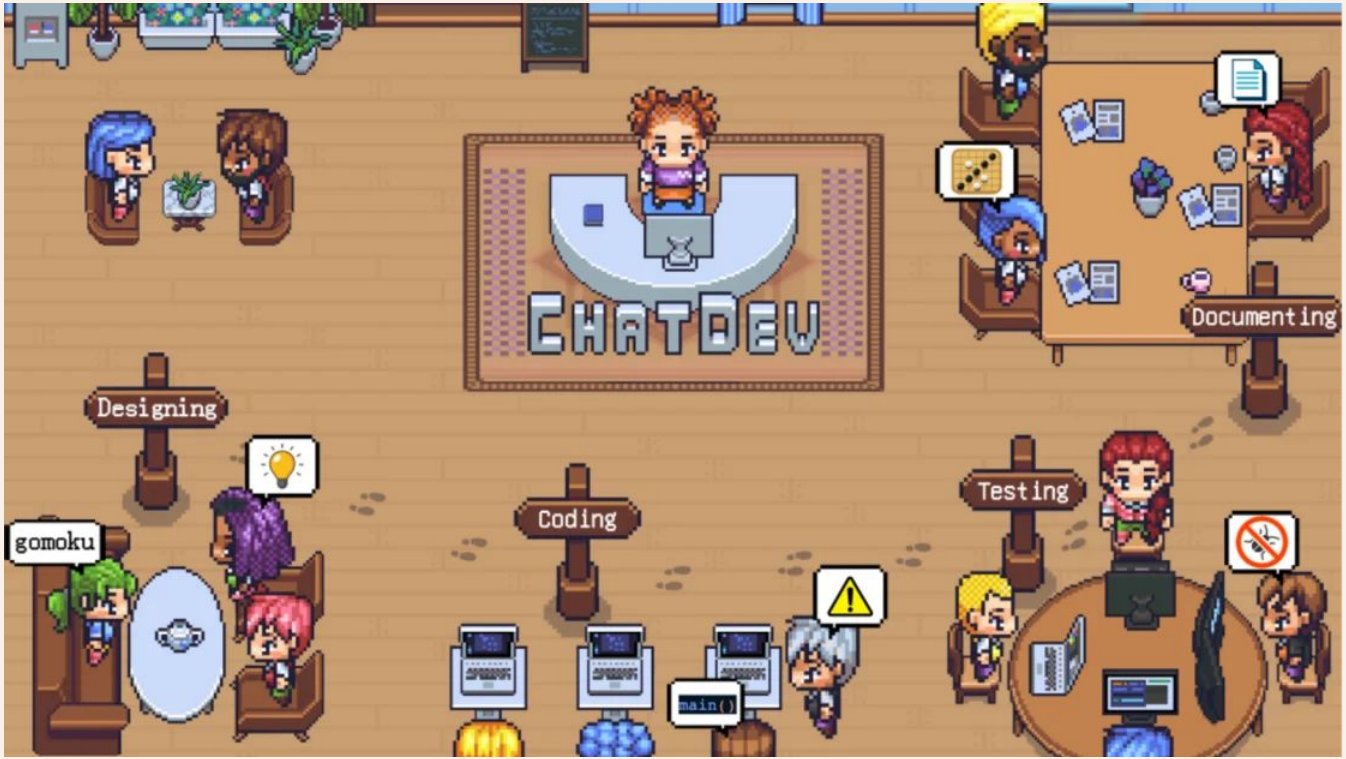


Ai2

Multi-agent Collab

Agents work together to solve a complex task.

- Autogen
- CrewAI
- **ChatDev**
- DataVoyager



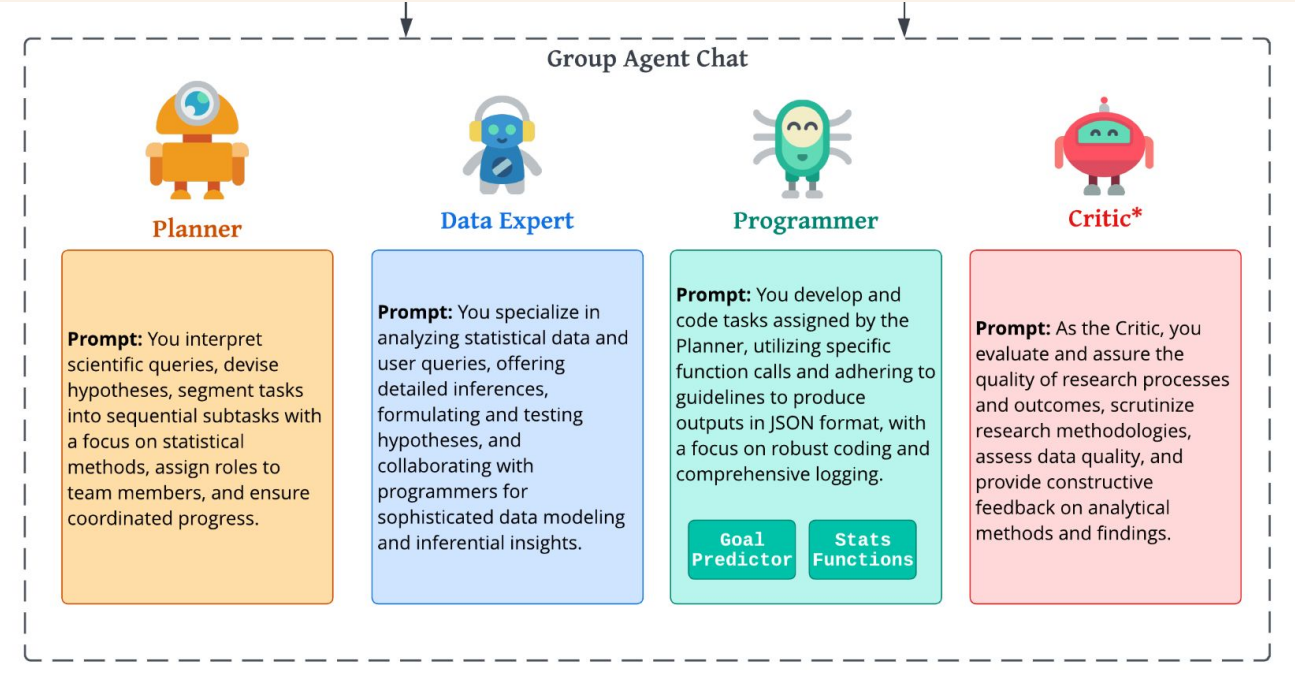
ChatDev: Communicative Agents for Software Development. *Qian et. al. ACL 2024*

Ai2

Multi-agent Collab

Agents work together to solve a complex task.

- Autogen
- CrewAI
- ChatDev
- DataVoyager
  - Agents
  - Memory
  - Functions
  - Literature



Data-driven Discovery with Large Generative Models. *Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, Peter Clark. ICML 2024.*

Ai2

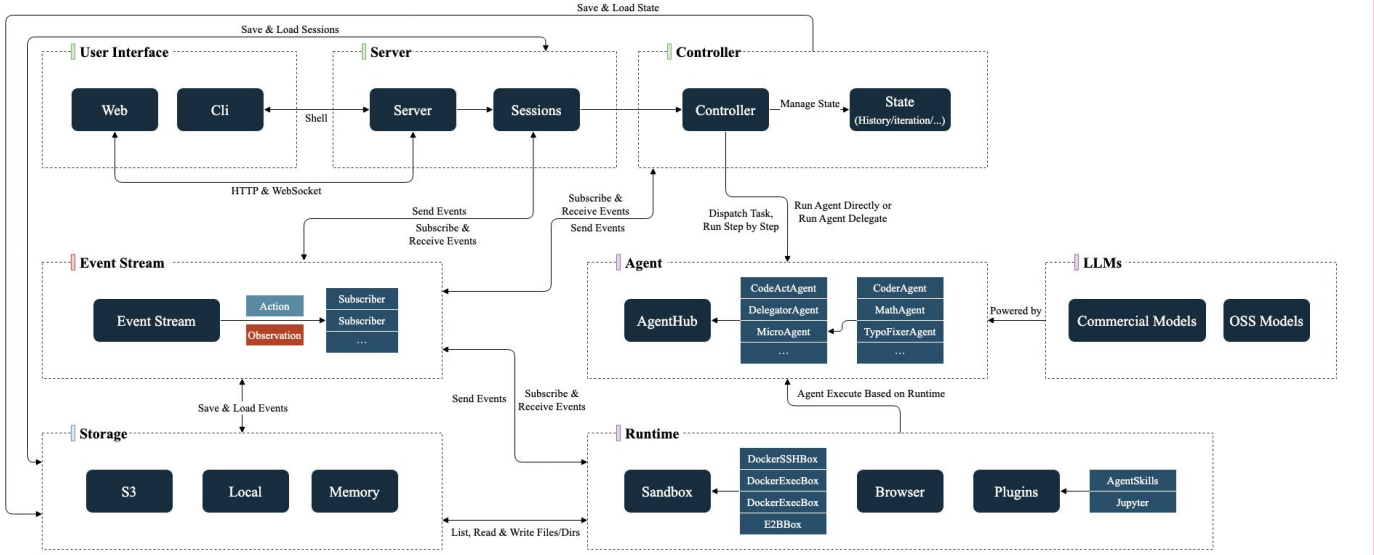
Ai2 UMASS AMHERST OpenLocus

Let’s go over 2 concrete examples!

Software Engg with LLMs

(not just code generation)

OpenHands by CMU & UIUC



Full coding env; **not just code gen**

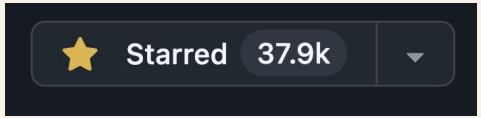
Access to agents

Access to runtime

Access to tools

Pluggable LLMs

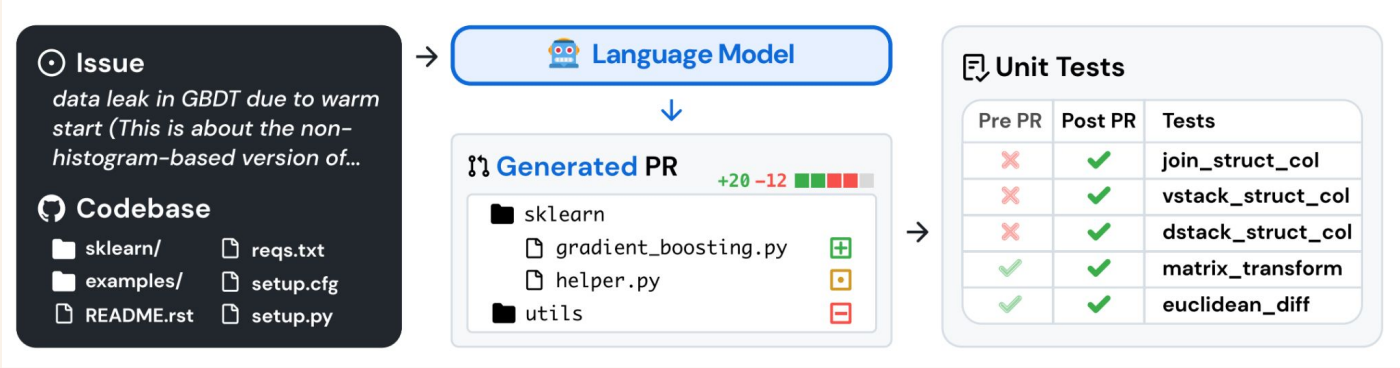
Top LLM repo on Github



Ai2

Carnegie Mellon University

SWE Bench by Princeton



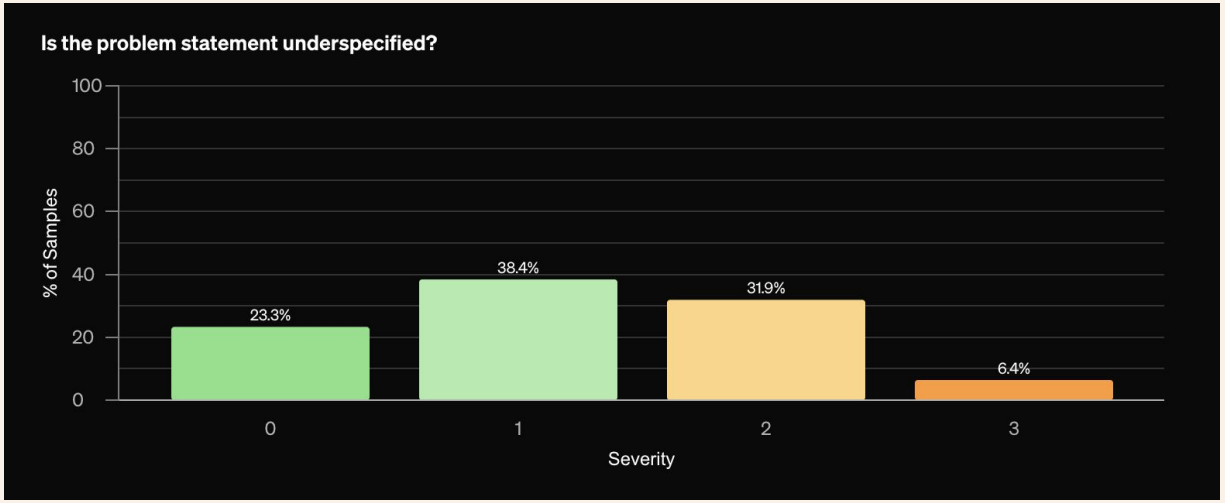
SWE-bench sources task instances from **real-world Python repositories** by connecting GitHub issues to merged pull request solutions that resolve related tests.

Provided with the **issue text and a codebase snapshot, models generate a patch** that is evaluated against real tests.

Can Language Models Resolve Real-World GitHub Issues? *Carlos E. Jimenez*, John Yang*, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, Karthik Narasimhan. ICLR 2024*

Ai2

SWE Bench Verified by OpenAI



SWE-bench Verified manually screened by 93 SWEs for 1,699 random samples.

Whether we consider the issue description to be **underspecified and hence unfair to be testing on.**

Whether the FAIL_TO_PASS unit tests **filter out valid solutions.**

https://openai.com/index/introducing-swe-bench-verified/

Neil Chowdhury, James Aung, Chan Jun Shern, Oliver Jaffe, Dane Sherburn, Giulio Starace, Evan Mays, Rachel Dias, Marwan Aljubeh, Mia Glaese, Carlos E. Jimenez, John Yang, Kevin Liu, Aleksander Madry

OpenAI

Ai2

OpenHands - CodeAct Agent

First agent to **cross 50%** in SWE-Bench Verified

Task planning by developing capabilities for bug detection, codebase management, and optimization

Made a number of fixes to make it easier for agents to **traverse directories**

Switched to use **function calling**, a method used by language models to more precisely specify the functions available to them

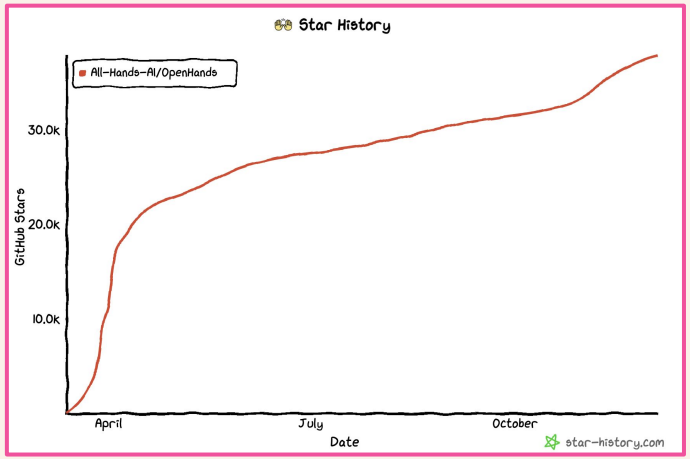
Agent + Model	Score
OpenHands + CodeAct v2.1 + Claude 3.5 Sonnet	**53.00**
Anthropic Tools + Claude 3.5 Sonnet	49.00
Anthropic Tools + Claude 3.5 Haiku	40.60
Composio SWEkit + Claude 3.5 Sonnet	40.60
SWE-agent + Claude 3.5 Sonnet	33.60
SWE-agent + Claude 3 Opus	18.20
RAG + Claude 3 Opus	**7.00**
RAG + Claude 2	4.40

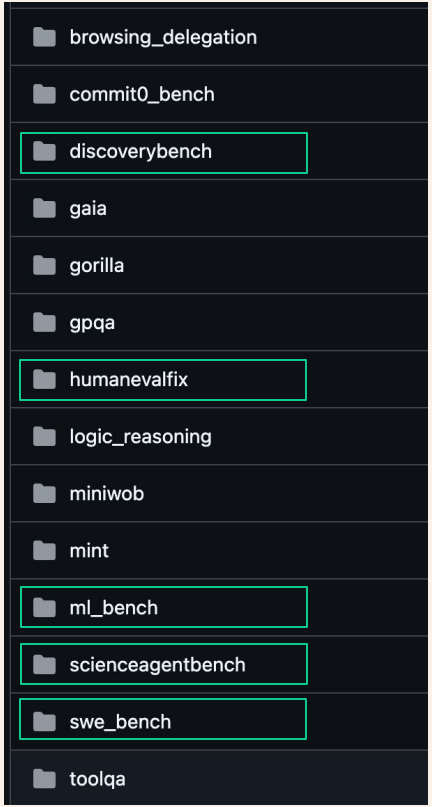
Contributing to OSS LLM Dev with OpenHands

We are contributing for **better evals & data science agents** - can discuss more offline.

Do try contributing to OpenHands. Their **aim is to have full replication of production-grade applications with LLMs**

https://github.com/All-Hands-AI/OpenHands/blob/main/CONTRIBUTING.md

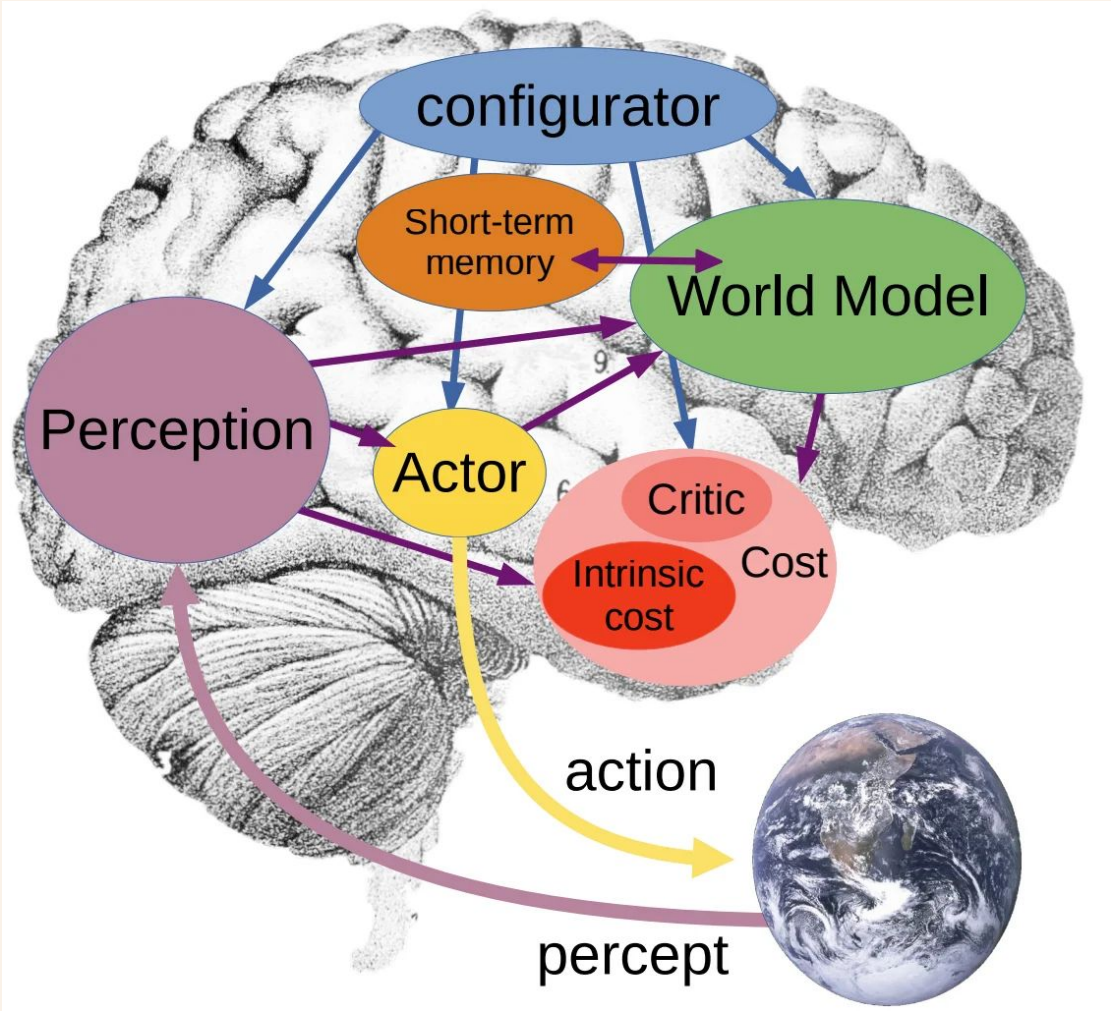




Ai2

V Jepa

By Meta/ Yann Le Cunn



Thank you!

Ai2