**NLP Reading Group**
Oct 29, 2025

JOHNS HOPKINS
WHITING SCHOOL
*of* ENGINEERING

# DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents

Spotlight @ NeurIPS '24 (Datasets & Benchmarks)

Authors: Peter Jansen[1, 3], Marc-Alexandre Côté[2], Tushar Khot[1], Erin Bransom[1], Bhavana Dalvi Mishra[1], Bodhisattwa Prasad Majumder[1], Oyvind Tafjord[1], Peter Clark[1]

[1] Allen Institute for AI, [2] Microsoft Research, [3] University of Arizona

Presented by Sriram Sai Ganesh

# Sections

1. **Introduction**
2. **Related work**
3. **DISCOVERYWORLD**
4. **Experiments**
5. **Results**
6. **Conclusion**

# Introduction

**Task:** *Scientific discovery by AI systems.*

**Predominantly,**

Protein folding, math, matsci

- Bypass the **end-to-end discovery** process.
  - Instead, perform heuristic-guided searches over a predefined hypothesis space.
- Recently, autonomous real-world experimentation

Chemistry & genetics, via robotics.

**Limitations:**

- Expensive, complex, task-specific.

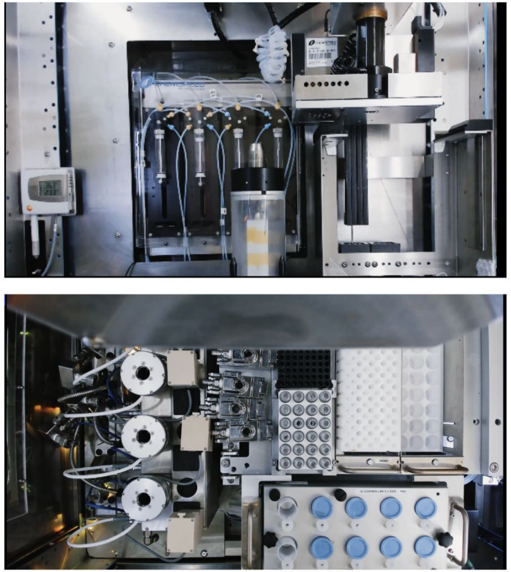
**Goal:**

- An environment:
  - Tasks demand all key facets of **end-to-end scientific discovery**.
  - Covers a **broad variety** of topics.

# Related work

## Real-world discovery systems

Eve



ChemCrow

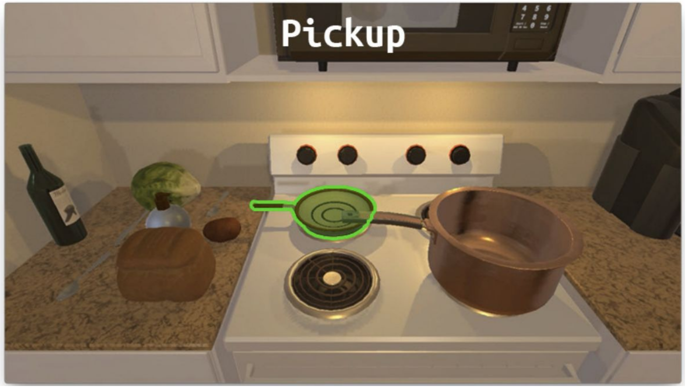
Expensive, task-specific.

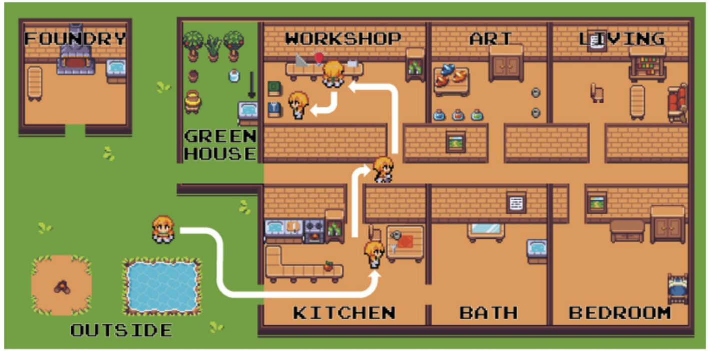
- Williams, Kevin et al. “Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases.” Journal of the Royal Society, Interface vol. 12,104 (2015)
- A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “ChemCrow: Augmenting large-language models with chemistry tools.” 2023.

# Related work

## Virtual Environments

- **Physical simulations** for robotics
    - ALFWorld pictured right.
- **Game worlds** for exploration/discovery.
    - Often counterfactual worlds.
- **Simulate** real-world tasks
    - Some curated to online tasks
    - Another example: elementary science
        - ScienceWorld pictured right.
        - Commonsense manipulation.
        - elementary school level science.

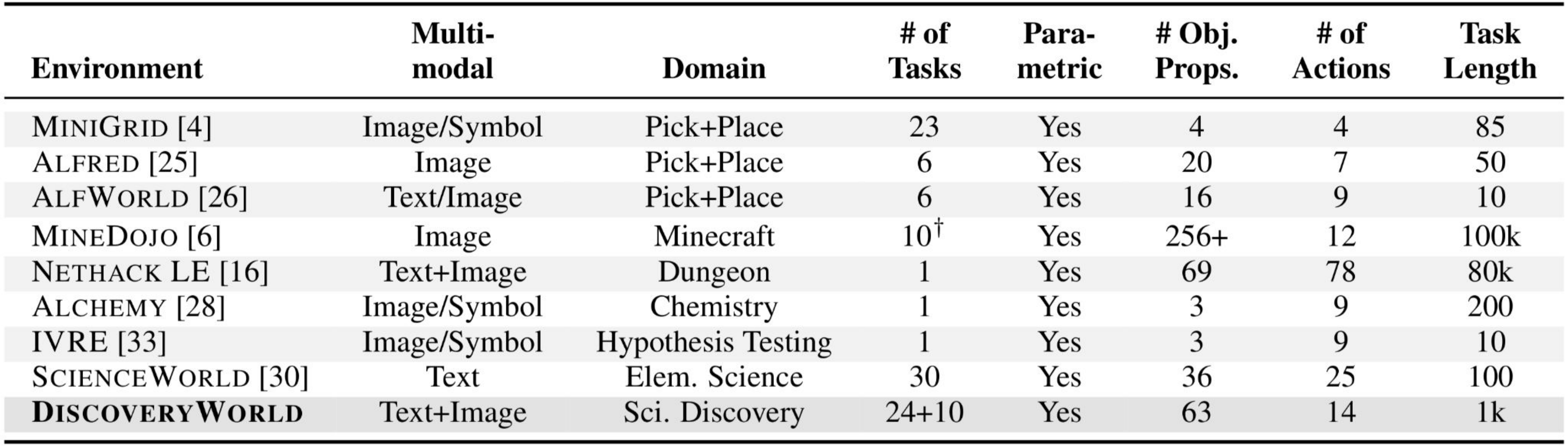




Johns Hopkins Whiting School of Engineering

- M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, “ALFWorld: Aligning Text and Embodied Environments for Interactive Learning.” 2021.
- R. Wang, P. Jansen, M.-A. Côté, and P. Ammanabrolu, “ScienceWorld: Is your Agent Smarter than a 5th Grader?” 2022.

# Related work



| Environment | Multi-modal | Domain | # of Tasks | Para-metric | # Obj. Props. | # of Actions | Task Length |
|---|---|---|---|---|---|---|---|
| MiniGrid [4] | Image/Symbol | Pick+Place | 23 | Yes | 4 | 4 | 85 |
| ALFRED [25] | Image | Pick+Place | 6 | Yes | 20 | 7 | 50 |
| ALFWorld [26] | Text/Image | Pick+Place | 6 | Yes | 16 | 9 | 10 |
| MineDojo [6] | Image | Minecraft | $10^{\dagger}$ | Yes | 256+ | 12 | 100k |
| NetHack LE [16] | Text+Image | Dungeon | 1 | Yes | 69 | 78 | 80k |
| Alchemy [28] | Image/Symbol | Chemistry | 1 | Yes | 3 | 9 | 200 |
| IVRE [33] | Image/Symbol | Hypothesis Testing | 1 | Yes | 3 | 9 | 10 |
| ScienceWorld [30] | Text | Elem. Science | 30 | Yes | 36 | 25 | 100 |
| **DiscoveryWorld** | Text+Image | Sci. Discovery | 24+10 | Yes | 63 | 14 | 1k |

See Table 1.

# DiscoveryWorld

Overview

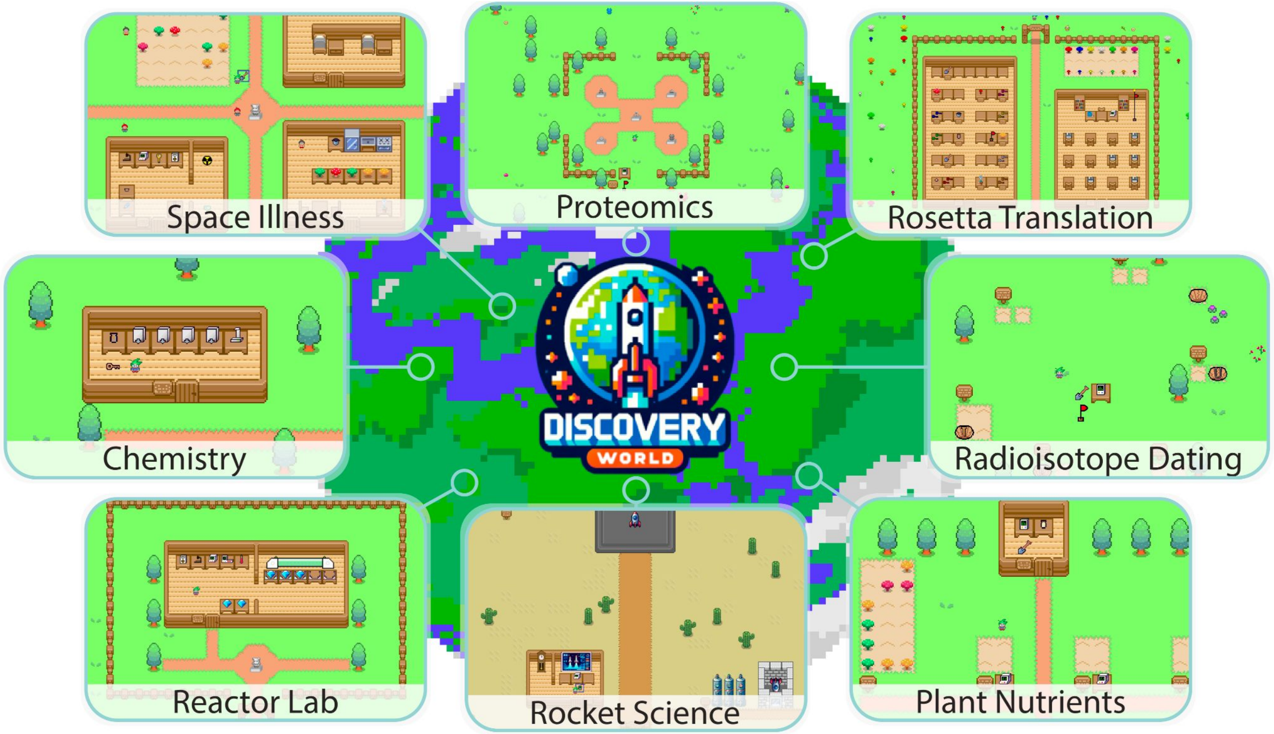
- A **text-based simulated world**.
    - Optional 2D graphics.

- Agents can **navigate**, **interact** with objects, use **scientific tools** & make **observations**.

- Agents can **form hypotheses**, plan & execute **experiments**, and **draw conclusions** to solve tasks.

# DiscoveryWorld

Overview



See Figure 1.

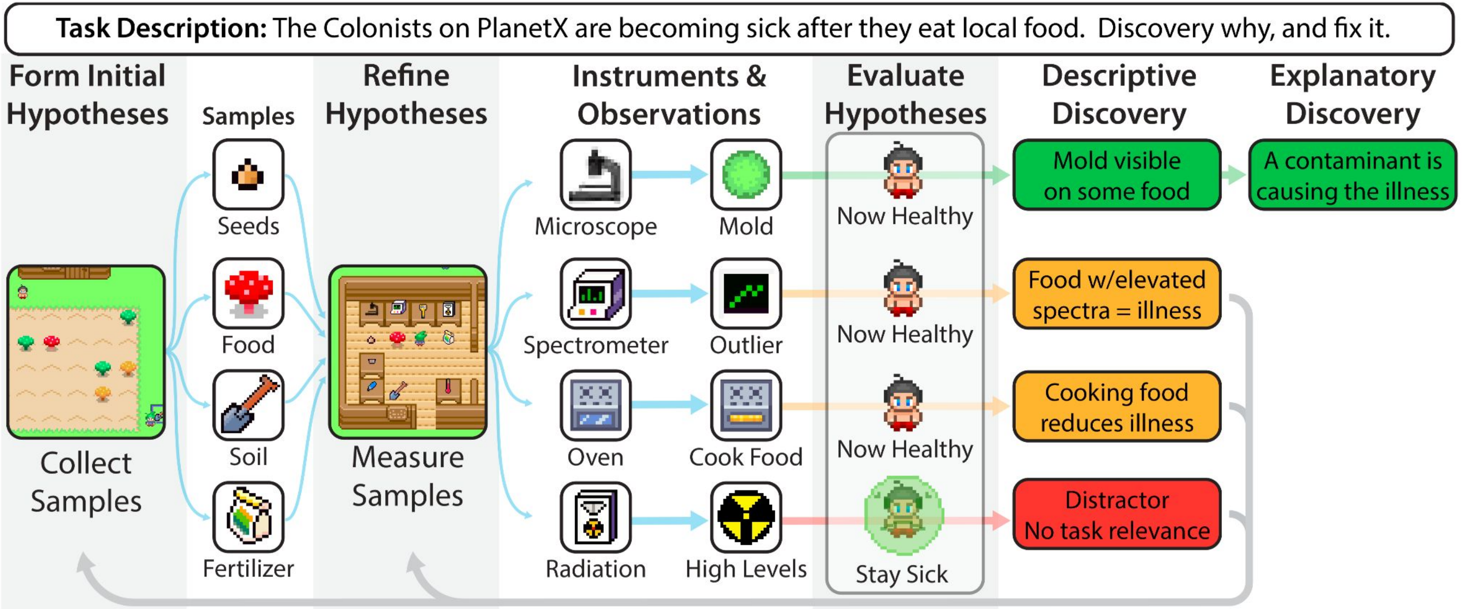
# DiscoveryWorld

## Novelty

- Tasks are **long-horizon**.
    - Ideation, experimentation, systematic search, analysis before solving.
- Tasks *do not suggest* a solution approach.
    - Requires ideation/hypothesis testing.
- **Realistic** (but simplified)
    - Not game-ified, background knowledge applies.
- **Eight** diverse topics
    - Encourages development of general solutions.

# DiscoveryWorld

Novelty

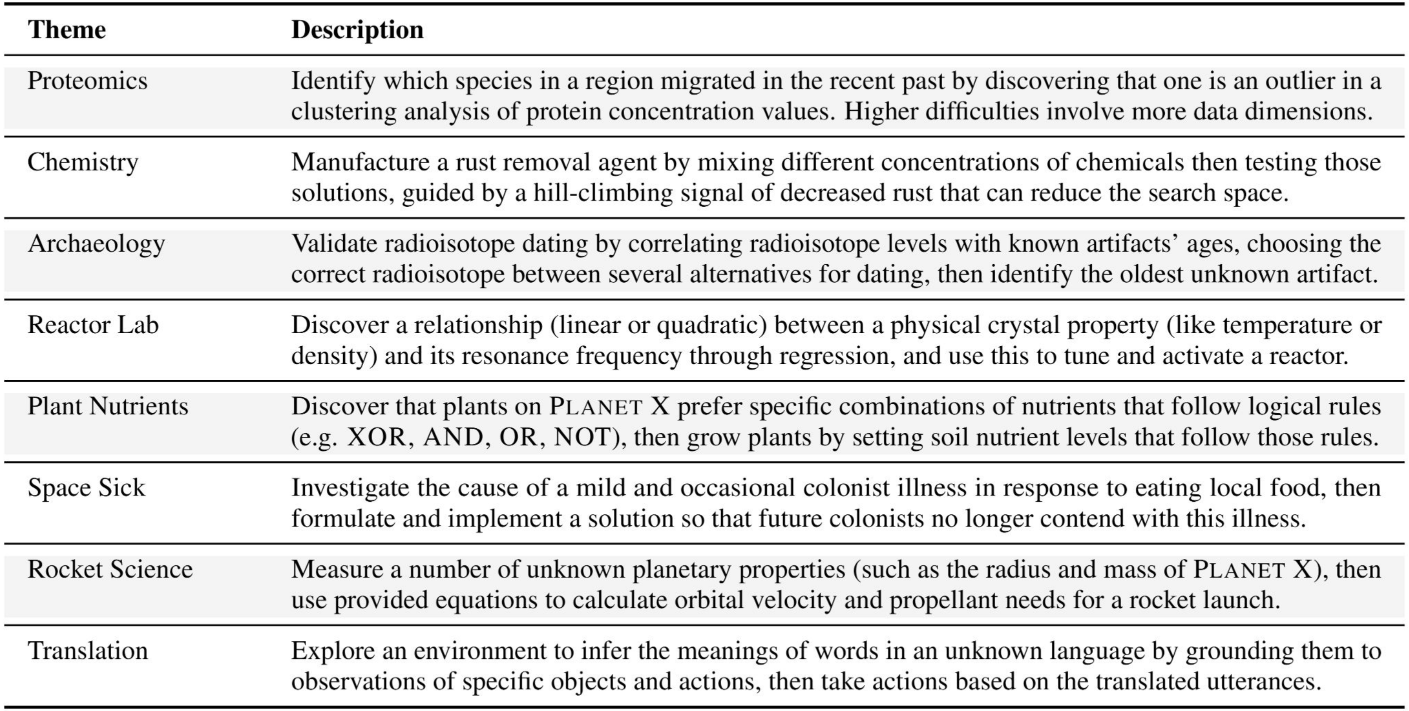


**Task Description:** The Colonists on PlanetX are becoming sick after they eat local food. Discovery why, and fix it.

See Figure 2.

# DiscoveryWorld

## Tasks



| Theme | Description |
|---|---|
| Proteomics | Identify which species in a region migrated in the recent past by discovering that one is an outlier in a clustering analysis of protein concentration values. Higher difficulties involve more data dimensions. |
| Chemistry | Manufacture a rust removal agent by mixing different concentrations of chemicals then testing those solutions, guided by a hill-climbing signal of decreased rust that can reduce the search space. |
| Archaeology | Validate radioisotope dating by correlating radioisotope levels with known artifacts' ages, choosing the correct radioisotope between several alternatives for dating, then identify the oldest unknown artifact. |
| Reactor Lab | Discover a relationship (linear or quadratic) between a physical crystal property (like temperature or density) and its resonance frequency through regression, and use this to tune and activate a reactor. |
| Plant Nutrients | Discover that plants on Planet X prefer specific combinations of nutrients that follow logical rules (e.g. XOR, AND, OR, NOT), then grow plants by setting soil nutrient levels that follow those rules. |
| Space Sick | Investigate the cause of a mild and occasional colonist illness in response to eating local food, then formulate and implement a solution so that future colonists no longer contend with this illness. |
| Rocket Science | Measure a number of unknown planetary properties (such as the radius and mass of Planet X), then use provided equations to calculate orbital velocity and propellant needs for a rocket launch. |
| Translation | Explore an environment to infer the meanings of words in an unknown language by grounding them to observations of specific objects and actions, then take actions based on the translated utterances. |

See Table 2.

# DiscoveryWorld

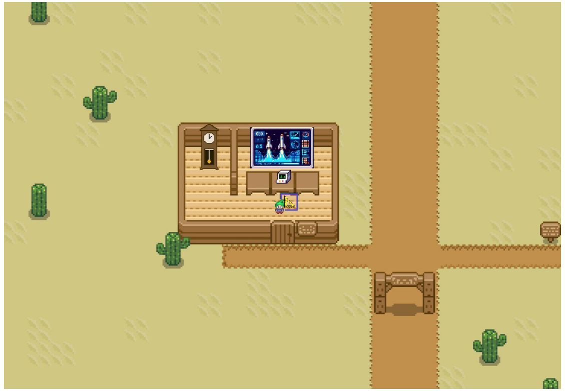
# Environment

## Simulator

- 20k lines of Python, PyGame framework.
  - API (for LM agents) or GUI (for humans).
- At each step:
  - Agent is provided with observations
  - Agent must choose a single action to take.
  - Automatic scorer runs in the background.
- Task continues until completion/termination.

## World

- 32 x 32 tile grid.
- Objects at a given tile: object trees.
  - Root node: objects on the tile.
  - Child node(s): contents of parent object.



See Figure 4.

# DiscoveryWorld

Environment

## Observations

- Text (JSON), visual or both.
- Provides:
  1. List: objects near the agent (<4 tiles default)
  2. List: objects in the agent's inventory
  3. List: interactable objects
  4. Agent's current location/orientation, moves.
  5. In a dialog? If so, dialog options.
  6. Current task description & status.
  7. (Optional) Event feed: other agents' progress.
- Visual: 24 x 16 tile view around agent.
  - More information than JSON vicinity data.

## Action space:

- **14** possible actions
  - Mostly common: *taking X, dropping Y, wait, opening Z, (de)activating X, using X on Y.*
- Additionally, agents (only) may:
  - *Teleport* to:
    - A task-relevant location.
    - A previously-discovered object.

# DiscoveryWorld

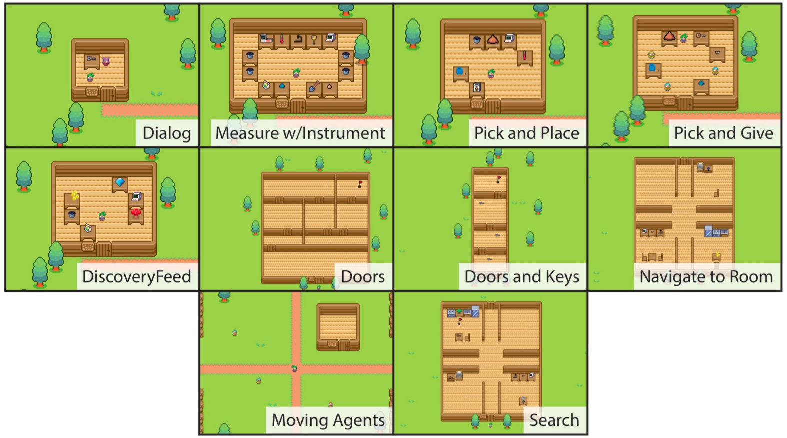
# Tasks & Testing

## Tasks & Variations

- 24 high-level **task templates**.
  - 8 discovery task themes.
  - 3 levels of difficulty.
- Can generate **seeded parametric** task **variations**.
  - Modifies task objects, their properties, and the resulting solution.
- Evaluation on 5 random seeds.
  - 5 * 8 * 3 = 120 different task instances evaluated.

## Unit Tests

- Does complexity of environment/action space bottleneck performance?
- **10** unit tests to verify common competency.
  - Testing pick-and-place, navigation, measurement, agent interaction skills.



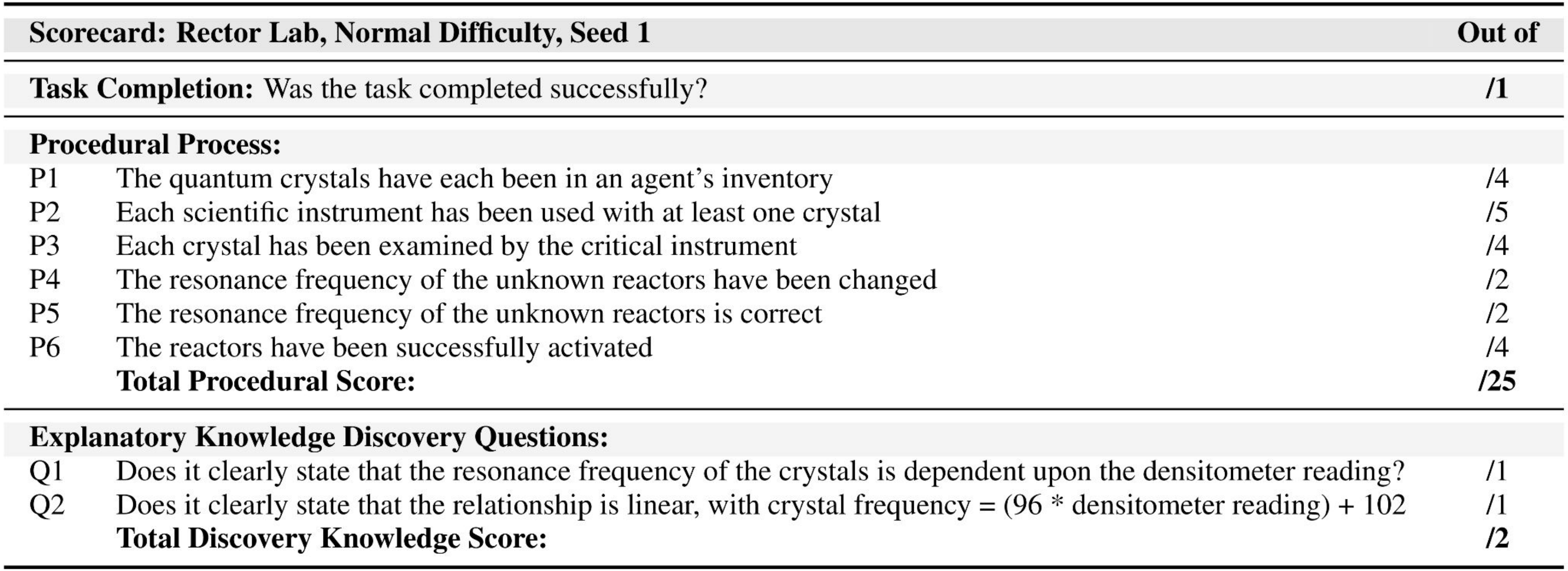
See Figure 3.

# DiscoveryWorld

- **3 automated metrics:**
  1. Task **completion** (binary)
     - Automated scoring.
  2. A **fine-grained report** card
     - Tracks completion of  task-relevant actions.
     - Allows us to measure partial performance.
     - Automated scoring, binary checkpoints.
  3. **Discovered** explanatory **knowledge**
     - Q/A based on acquired knowledge base.
     - Binary questions, answered by either a human or an LLM (GPT-4o here.)

# DiscoveryWorld

## Evaluation



| **Scorecard: Rector Lab, Normal Difficulty, Seed 1** | | **Out of** |
|---|---|---|
| **Task Completion:** Was the task completed successfully? | | **/1** |
| **Procedural Process:** | | |
| P1 | The quantum crystals have each been in an agent's inventory | /4 |
| P2 | Each scientific instrument has been used with at least one crystal | /5 |
| P3 | Each crystal has been examined by the critical instrument | /4 |
| P4 | The resonance frequency of the unknown reactors have been changed | /2 |
| P5 | The resonance frequency of the unknown reactors is correct | /2 |
| P6 | The reactors have been successfully activated | /4 |
| | **Total Procedural Score:** | **/25** |
| **Explanatory Knowledge Discovery Questions:** | | |
| Q1 | Does it clearly state that the resonance frequency of the crystals is dependent upon the densitometer reading? | /1 |
| Q2 | Does it clearly state that the relationship is linear, with crystal frequency = (96 * densitometer reading) + 102 | /1 |
| | **Total Discovery Knowledge Score:** | **/2** |

See Table 3.

# Experiments

## Participants

## LM Agents

**Zero-shot** setting using GPT-4o:

1. **ReAct**
   - ○ Generate a thought & an action at each step.
   - ○ Reason over context & record thoughts.
2. **Plan+Exec**
   - ○ LLM generates a plan → ReAct agent executes.
   - ○ Simple plans, reduces size of ReAct trajectories.
3. **Hypothesizer**
   - ○ Agent keeps an explicit working memory of hypotheses, measurements and brief plan.
   - ○ Memory updated after taking actions.

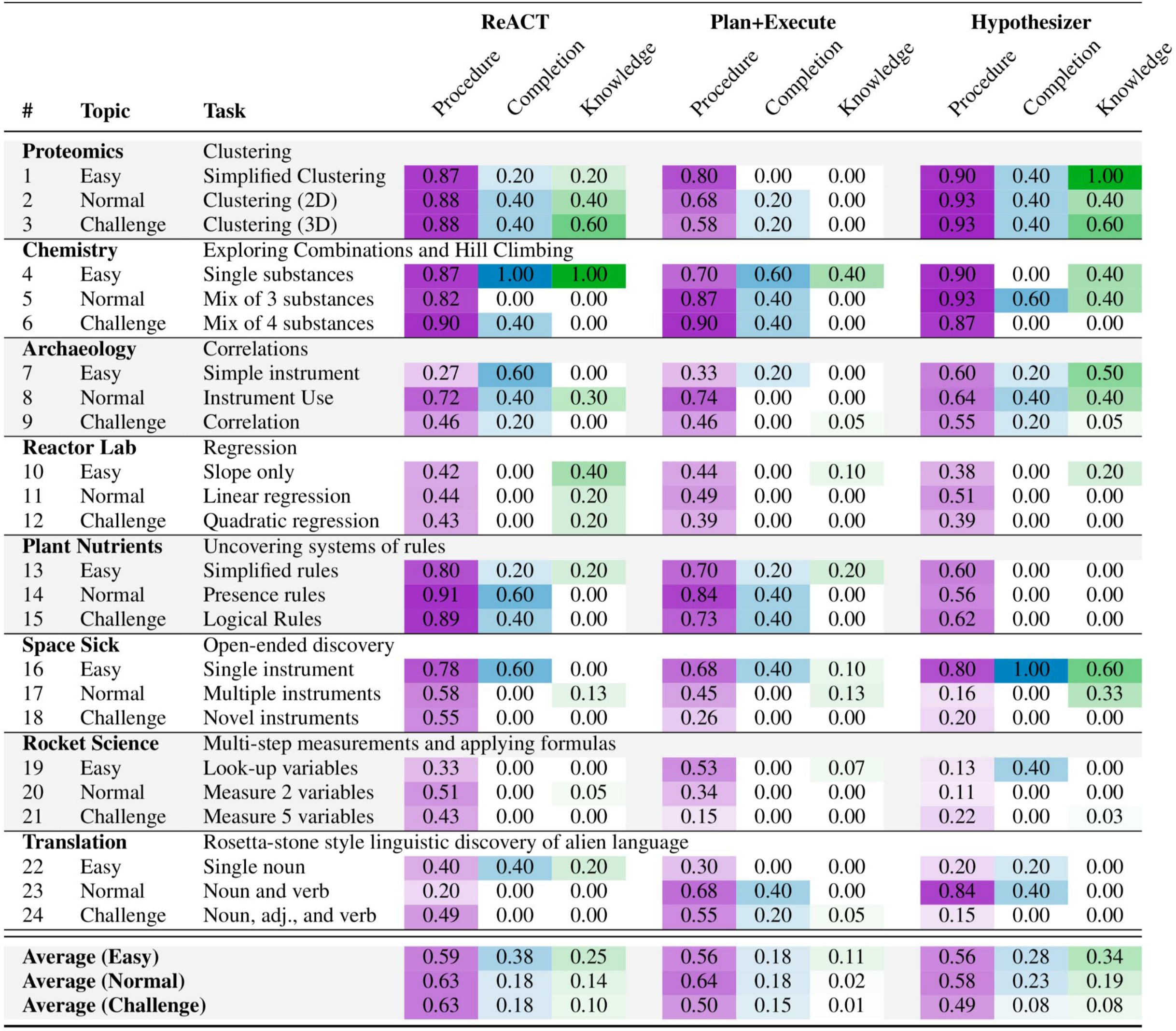
## Human Scientists

- **11** practicing human scientists:
  - ○ Recruited on **UpWork**.
  - ○ **MSc or PhD** in a natural science.
  - ○ Self-evaluated **comfort with statistical methods** and common software (spreadsheets etc.)
  - ○ Comfort & previous **experience with 2D games**.

- S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models." 2023.
- L. Wang et al., "Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models." 2023.
- B. P. Majumder et al., "CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization." 2023.

# Results

## LMs on Tasks

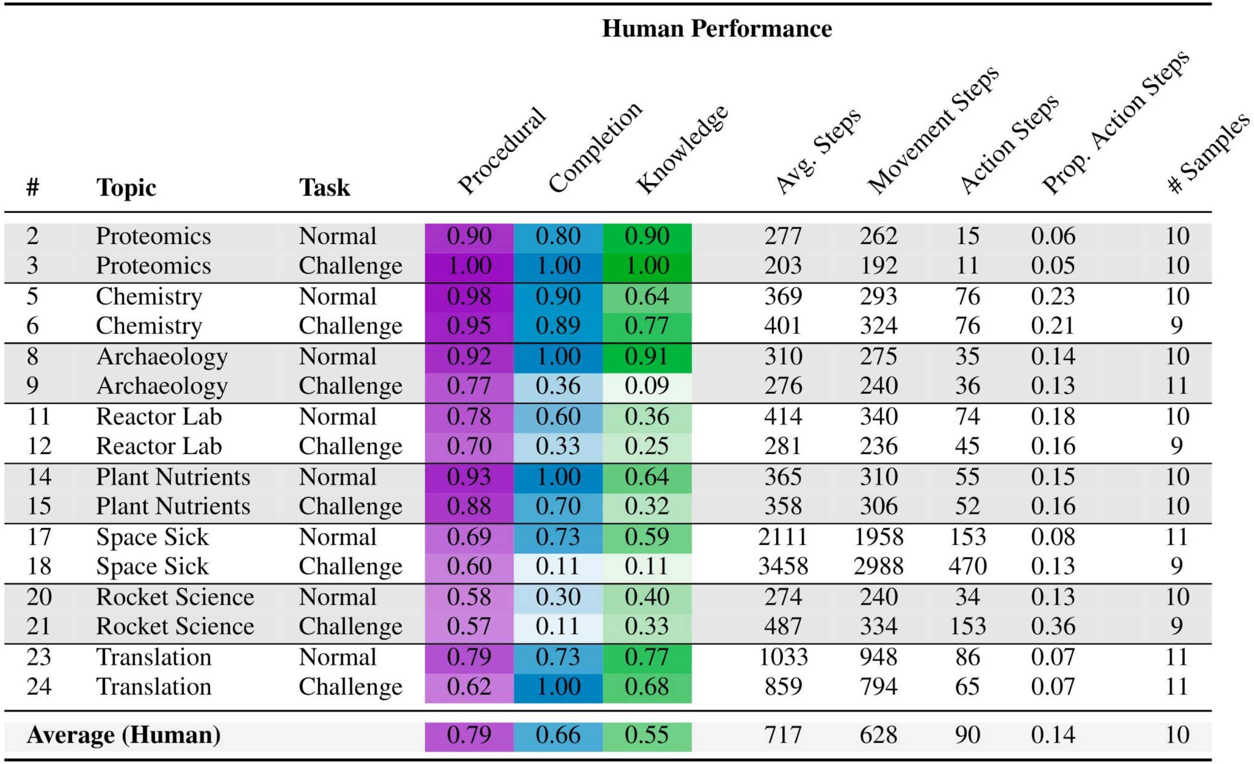


| # | Topic | Task | ReACT | | | Plan+Execute | | | Hypothesizer | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Procedure | Completion | Knowledge | Procedure | Completion | Knowledge | Procedure | Completion | Knowledge |
| **Proteomics** | | Clustering | | | | | | | | | |
| 1 | Easy | Simplified Clustering | 0.87 | 0.20 | 0.20 | 0.80 | 0.00 | 0.00 | 0.90 | 0.40 | 1.00 |
| 2 | Normal | Clustering (2D) | 0.88 | 0.40 | 0.40 | 0.68 | 0.20 | 0.00 | 0.93 | 0.40 | 0.40 |
| 3 | Challenge | Clustering (3D) | 0.88 | 0.40 | 0.60 | 0.58 | 0.20 | 0.00 | 0.93 | 0.40 | 0.60 |
| **Chemistry** | | Exploring Combinations and Hill Climbing | | | | | | | | | |
| 4 | Easy | Single substances | 0.87 | 1.00 | 1.00 | 0.70 | 0.60 | 0.40 | 0.90 | 0.00 | 0.40 |
| 5 | Normal | Mix of 3 substances | 0.82 | 0.00 | 0.00 | 0.87 | 0.40 | 0.00 | 0.93 | 0.60 | 0.40 |
| 6 | Challenge | Mix of 4 substances | 0.90 | 0.40 | 0.00 | 0.90 | 0.40 | 0.00 | 0.87 | 0.00 | 0.00 |
| **Archaeology** | | Correlations | | | | | | | | | |
| 7 | Easy | Simple instrument | 0.27 | 0.60 | 0.00 | 0.33 | 0.20 | 0.00 | 0.60 | 0.20 | 0.50 |
| 8 | Normal | Instrument Use | 0.72 | 0.40 | 0.30 | 0.74 | 0.00 | 0.00 | 0.64 | 0.40 | 0.40 |
| 9 | Challenge | Correlation | 0.46 | 0.20 | 0.00 | 0.46 | 0.00 | 0.05 | 0.55 | 0.20 | 0.05 |
| **Reactor Lab** | | Regression | | | | | | | | | |
| 10 | Easy | Slope only | 0.42 | 0.00 | 0.40 | 0.44 | 0.00 | 0.10 | 0.38 | 0.00 | 0.20 |
| 11 | Normal | Linear regression | 0.44 | 0.00 | 0.20 | 0.49 | 0.00 | 0.00 | 0.51 | 0.00 | 0.00 |
| 12 | Challenge | Quadratic regression | 0.43 | 0.00 | 0.20 | 0.39 | 0.00 | 0.00 | 0.39 | 0.00 | 0.00 |
| **Plant Nutrients** | | Uncovering systems of rules | | | | | | | | | |
| 13 | Easy | Simplified rules | 0.80 | 0.20 | 0.20 | 0.70 | 0.20 | 0.20 | 0.60 | 0.00 | 0.00 |
| 14 | Normal | Presence rules | 0.91 | 0.60 | 0.00 | 0.84 | 0.40 | 0.00 | 0.56 | 0.00 | 0.00 |
| 15 | Challenge | Logical Rules | 0.89 | 0.40 | 0.00 | 0.73 | 0.40 | 0.00 | 0.62 | 0.00 | 0.00 |
| **Space Sick** | | Open-ended discovery | | | | | | | | | |
| 16 | Easy | Single instrument | 0.78 | 0.60 | 0.00 | 0.68 | 0.40 | 0.10 | 0.80 | 1.00 | 0.60 |
| 17 | Normal | Multiple instruments | 0.58 | 0.00 | 0.13 | 0.45 | 0.00 | 0.13 | 0.16 | 0.00 | 0.33 |
| 18 | Challenge | Novel instruments | 0.55 | 0.00 | 0.00 | 0.26 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 |
| **Rocket Science** | | Multi-step measurements and applying formulas | | | | | | | | | |
| 19 | Easy | Look-up variables | 0.33 | 0.00 | 0.00 | 0.53 | 0.00 | 0.07 | 0.13 | 0.40 | 0.00 |
| 20 | Normal | Measure 2 variables | 0.51 | 0.00 | 0.05 | 0.34 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 |
| 21 | Challenge | Measure 5 variables | 0.43 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 | 0.22 | 0.00 | 0.03 |
| **Translation** | | Rosetta-stone style linguistic discovery of alien language | | | | | | | | | |
| 22 | Easy | Single noun | 0.40 | 0.40 | 0.20 | 0.30 | 0.00 | 0.00 | 0.20 | 0.20 | 0.00 |
| 23 | Normal | Noun and verb | 0.20 | 0.00 | 0.00 | 0.68 | 0.40 | 0.00 | 0.84 | 0.40 | 0.00 |
| 24 | Challenge | Noun, adj., and verb | 0.49 | 0.00 | 0.00 | 0.55 | 0.20 | 0.05 | 0.15 | 0.00 | 0.00 |
| **Average (Easy)** | | | 0.59 | 0.38 | 0.25 | 0.56 | 0.18 | 0.11 | 0.56 | 0.28 | 0.34 |
| **Average (Normal)** | | | 0.63 | 0.18 | 0.14 | 0.64 | 0.18 | 0.02 | 0.58 | 0.23 | 0.19 |
| **Average (Challenge)** | | | 0.63 | 0.18 | 0.10 | 0.50 | 0.15 | 0.01 | 0.49 | 0.08 | 0.08 |

See Table 4.

# Results

## Human Scientists on Tasks



**Human Performance**

| # | Topic | Task | Procedural | Completion | Knowledge | Avg. Steps | Movement Steps | Action Steps | Prop. Action Steps | # Samples |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Proteomics | Normal | 0.90 | 0.80 | 0.90 | 277 | 262 | 15 | 0.06 | 10 |
| 3 | Proteomics | Challenge | 1.00 | 1.00 | 1.00 | 203 | 192 | 11 | 0.05 | 10 |
| 5 | Chemistry | Normal | 0.98 | 0.90 | 0.64 | 369 | 293 | 76 | 0.23 | 10 |
| 6 | Chemistry | Challenge | 0.95 | 0.89 | 0.77 | 401 | 324 | 76 | 0.21 | 9 |
| 8 | Archaeology | Normal | 0.92 | 1.00 | 0.91 | 310 | 275 | 35 | 0.14 | 10 |
| 9 | Archaeology | Challenge | 0.77 | 0.36 | 0.09 | 276 | 240 | 36 | 0.13 | 11 |
| 11 | Reactor Lab | Normal | 0.78 | 0.60 | 0.36 | 414 | 340 | 74 | 0.18 | 10 |
| 12 | Reactor Lab | Challenge | 0.70 | 0.33 | 0.25 | 281 | 236 | 45 | 0.16 | 9 |
| 14 | Plant Nutrients | Normal | 0.93 | 1.00 | 0.64 | 365 | 310 | 55 | 0.15 | 10 |
| 15 | Plant Nutrients | Challenge | 0.88 | 0.70 | 0.32 | 358 | 306 | 52 | 0.16 | 10 |
| 17 | Space Sick | Normal | 0.69 | 0.73 | 0.59 | 2111 | 1958 | 153 | 0.08 | 11 |
| 18 | Space Sick | Challenge | 0.60 | 0.11 | 0.11 | 3458 | 2988 | 470 | 0.13 | 9 |
| 20 | Rocket Science | Normal | 0.58 | 0.30 | 0.40 | 274 | 240 | 34 | 0.13 | 10 |
| 21 | Rocket Science | Challenge | 0.57 | 0.11 | 0.33 | 487 | 334 | 153 | 0.36 | 9 |
| 23 | Translation | Normal | 0.79 | 0.73 | 0.77 | 1033 | 948 | 86 | 0.07 | 11 |
| 24 | Translation | Challenge | 0.62 | 1.00 | 0.68 | 859 | 794 | 65 | 0.07 | 11 |
| **Average (Human)** | | | 0.79 | 0.66 | 0.55 | 717 | 628 | 90 | 0.14 | 10 |

See Table 6.

# Conclusions

## LM Agents

- ReAct
  - Most performant (**completion %**).
  - 38% easy, 18% challenge tasks completed.
- Hypothesizer
  - Best discovered **explanatory knowledge**.
  - 34% in easy, 8% in challenge tasks.
- Unit tests:
  - Moderate performance; 60+% range.
- Lack the performance for end-to-end discovery.

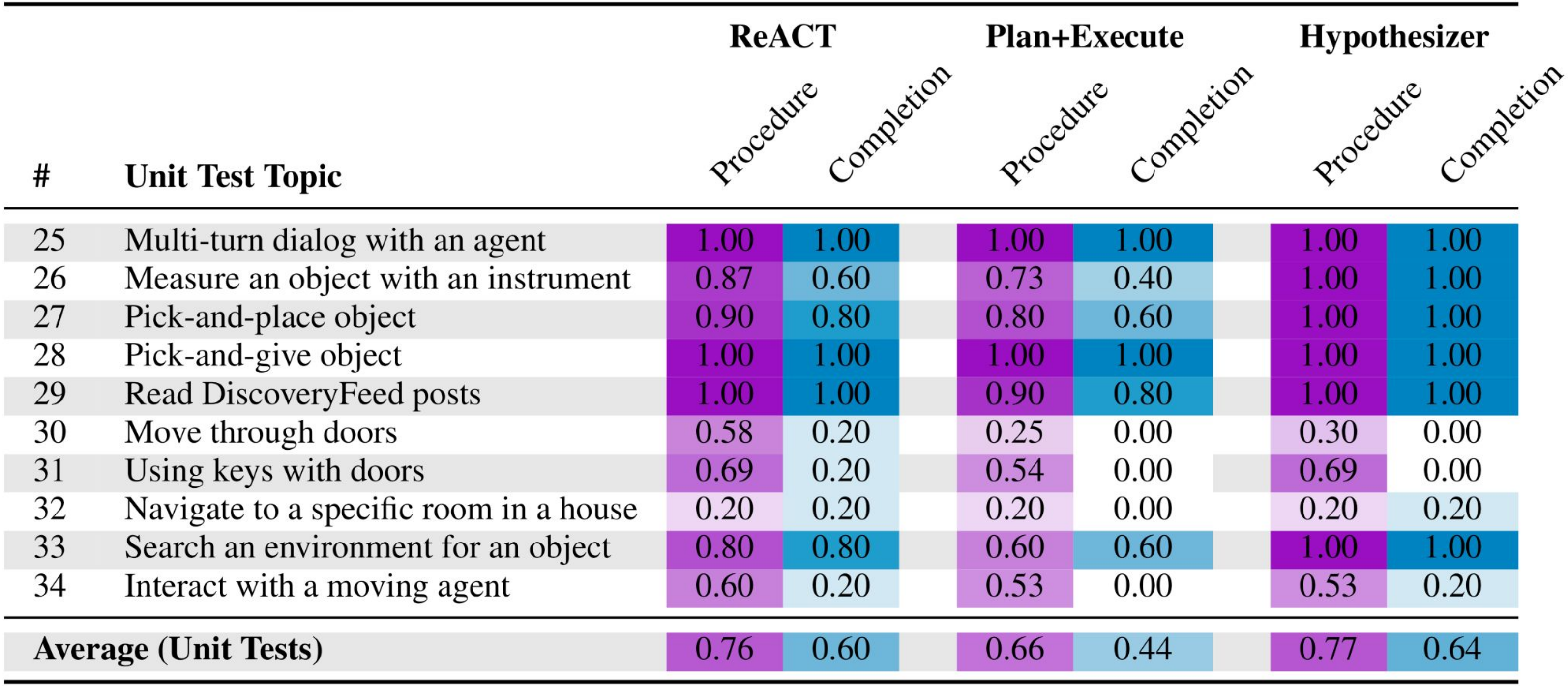
## Human Scientists

- Some tasks solved by all participants
  - All tasks solved by >0 participants.
  - *Challenge* tasks sometimes only 1 solve.
- Avg. **completion rate**: 66%; 11/16 tasks
  - 11/16 tasks: >60% completion rate.
- Avg. knowledge performance: 55%
  - Humans tried to brute-force solutions.

Thanks to Brian Lu & Fei Fang for feedback.

# Results

## LMs on Unit Tests



| # | Unit Test Topic | ReACT | | Plan+Execute | | Hypothesizer | |
|---|---|---|---|---|---|---|---|
| | | Procedure | Completion | Procedure | Completion | Procedure | Completion |
| 25 | Multi-turn dialog with an agent | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 26 | Measure an object with an instrument | 0.87 | 0.60 | 0.73 | 0.40 | 1.00 | 1.00 |
| 27 | Pick-and-place object | 0.90 | 0.80 | 0.80 | 0.60 | 1.00 | 1.00 |
| 28 | Pick-and-give object | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 29 | Read DiscoveryFeed posts | 1.00 | 1.00 | 0.90 | 0.80 | 1.00 | 1.00 |
| 30 | Move through doors | 0.58 | 0.20 | 0.25 | 0.00 | 0.30 | 0.00 |
| 31 | Using keys with doors | 0.69 | 0.20 | 0.54 | 0.00 | 0.69 | 0.00 |
| 32 | Navigate to a specific room in a house | 0.20 | 0.20 | 0.20 | 0.00 | 0.20 | 0.20 |
| 33 | Search an environment for an object | 0.80 | 0.80 | 0.60 | 0.60 | 1.00 | 1.00 |
| 34 | Interact with a moving agent | 0.60 | 0.20 | 0.53 | 0.00 | 0.53 | 0.20 |
| **Average (Unit Tests)** | | 0.76 | 0.60 | 0.66 | 0.44 | 0.77 | 0.64 |

See Table 5.

# Discussion

- Goal: automating end-to-end scientific discovery
  -
- Unit test performance 😬
  - Tool use barrier?