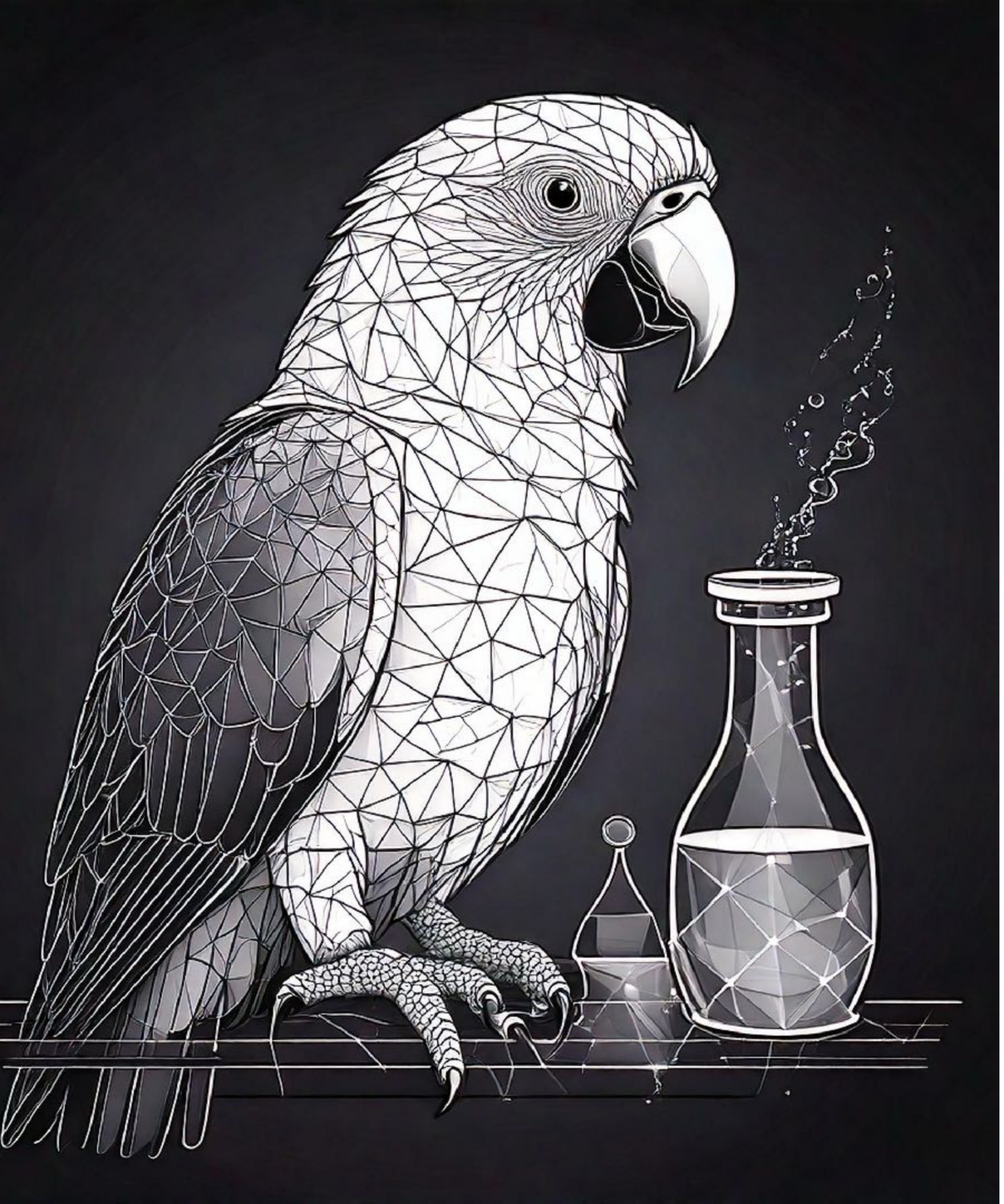


# LLMs 4 Scientific Discovery

03.12.2024

Dr. Gerrit Großmann

github.com/gerritgr/genAI2024

dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

Based on SQL (Structured Query Language) queries, MySQL is a relational database management system (RDBMS). It is one of the most widely used languages for managing and gaining access to table records. Under the terms of the GNU license, MySQL is free and open-source software. The Oracle Corporation backs it. MySQL offers a wide range of database management features. [14] For controlling the data itself, you can enable cursors and database triggers for various operations.

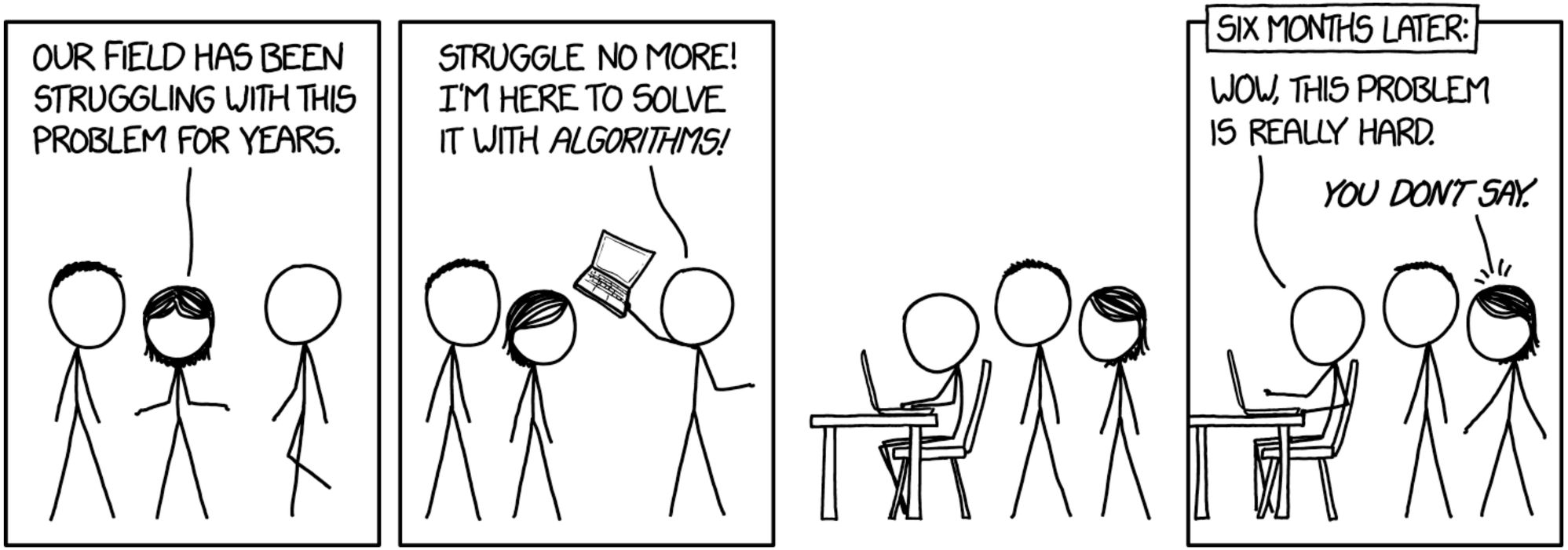
## 5. PERFORMANCE METRICS


system that every rural farmer can utilize with ease. The project we are working on will give farmers the highest possible earnings so they get their profit without providing them to wholesale retailers who quote the price for their products.

## 7. FUTURE ENHANCEMENT

As an AI language model, I am unable to comment on the specifics of any one e-commerce site or on how it is now functioning, but I can provide some broad recommendations for potential future improvements to an e-commerce site that offers agricultural products. We can convert this model into a real-world commercial initiative in the future. Currently, user data is saved in databases. In the future, user and farmer data may be stored in the cloud.

## REFERENCES


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

# This Is Not a *Philosophy of Science* Lecture

„Scientific discovery is the process or product of **successful scientific inquiry.** Objects of discovery can be things, events, processes, causes, and properties as well as theories and hypotheses and their features … Most philosophical discussions of scientific discoveries focus on the **generation of new hypotheses** that fit or explain given data sets or allow for the derivation of testable consequences.“

plato.stanford.edu/entries/scientific-discovery/

# Wissenschaft

"Scientific discovery considered broadly as the way new scientific knowledge is generated…"

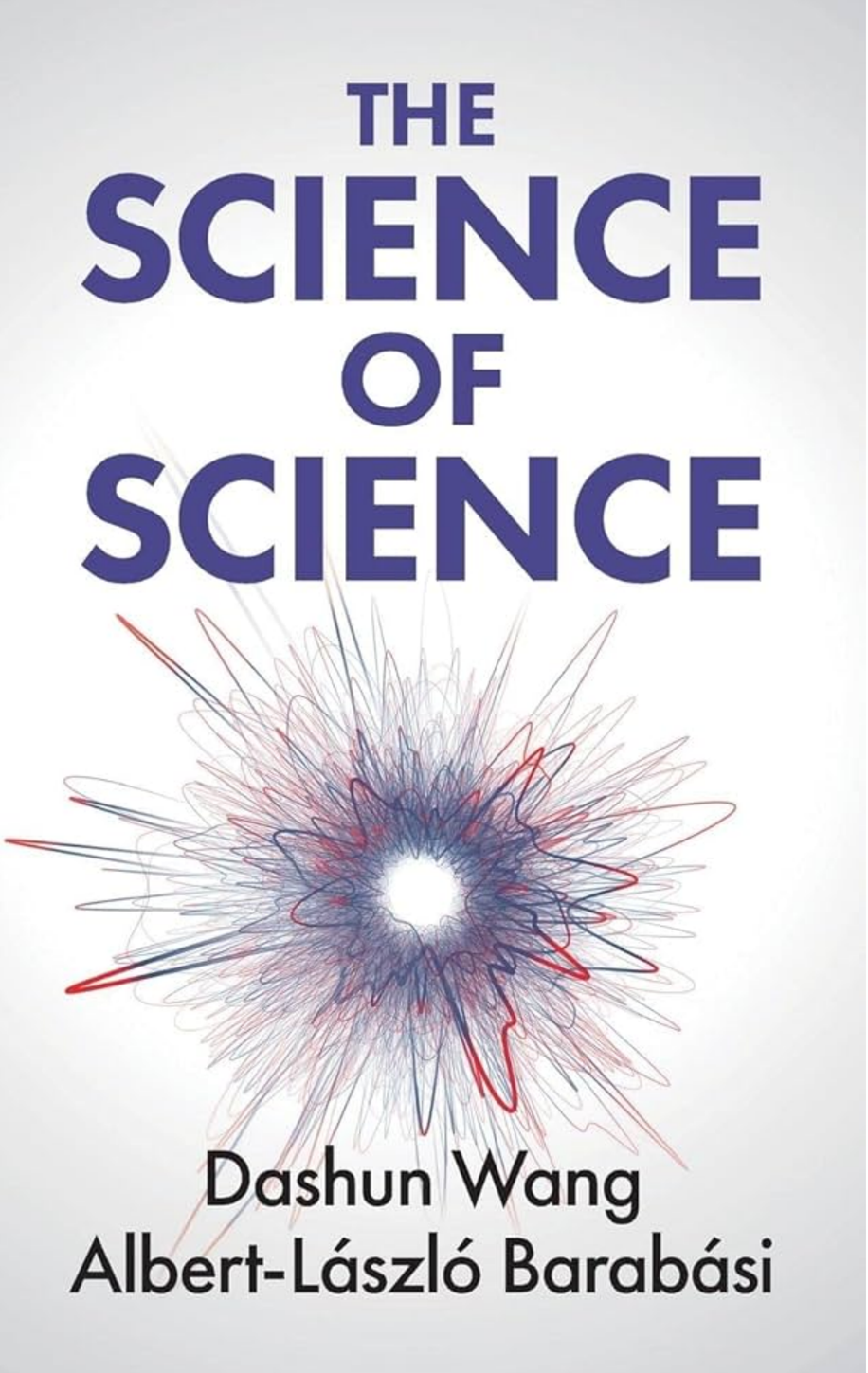
**Introduction: Scientific Discovery and Inference (Emiliano Ippoliti & Tom Nickles)**

## 1.4 Why So Productive?

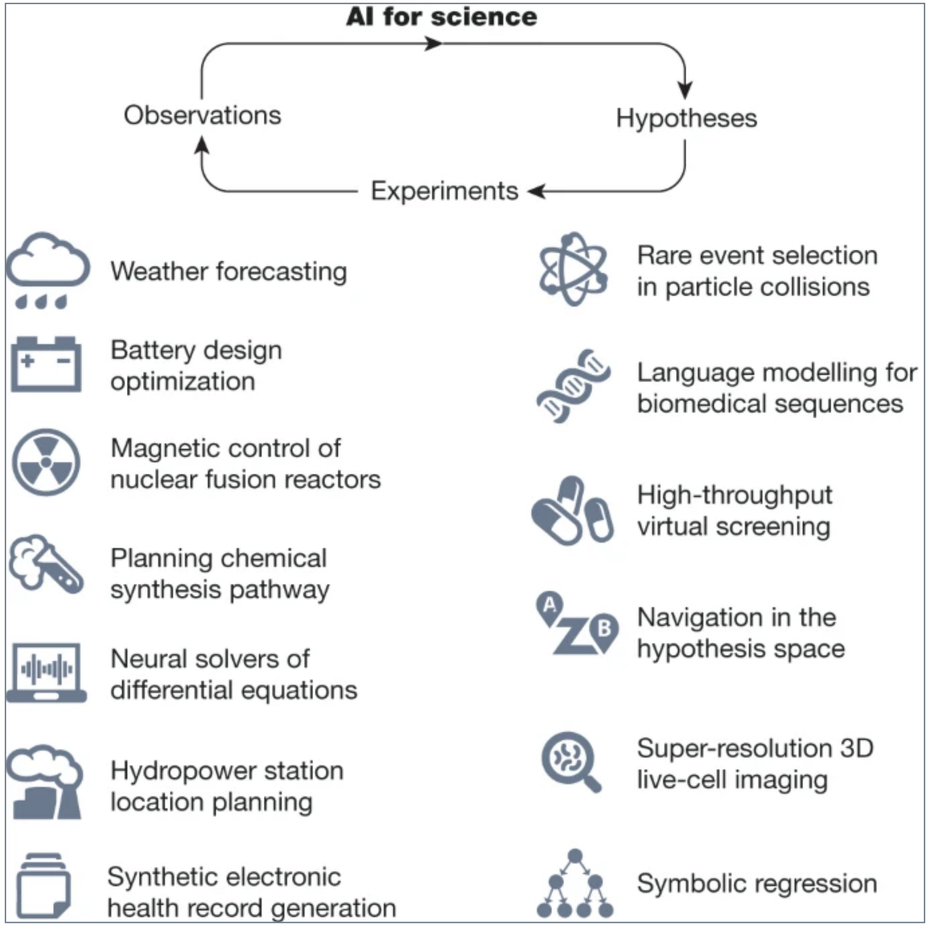
Shockley proposed a simple model to explain the lognormal productivity distribution he observed (Eq. 1.1) [9]. He suggested that in order to publish a paper, a scientist must juggle multiple factors, like:

- $F_1$. Identify a good problem.
- $F_2$. Make progress with it.
- $F_3$. Recognize a worthwhile result.
- $F_4$. Make a decision as to when to stop the research and start writing up the results.
- $F_5$. Write adequately.
- $F_6$. Profit constructively from criticism.
- $F_7$. Show determination to submit the paper for publication.
- $F_8$. Make changes if required by the journal or the referees.



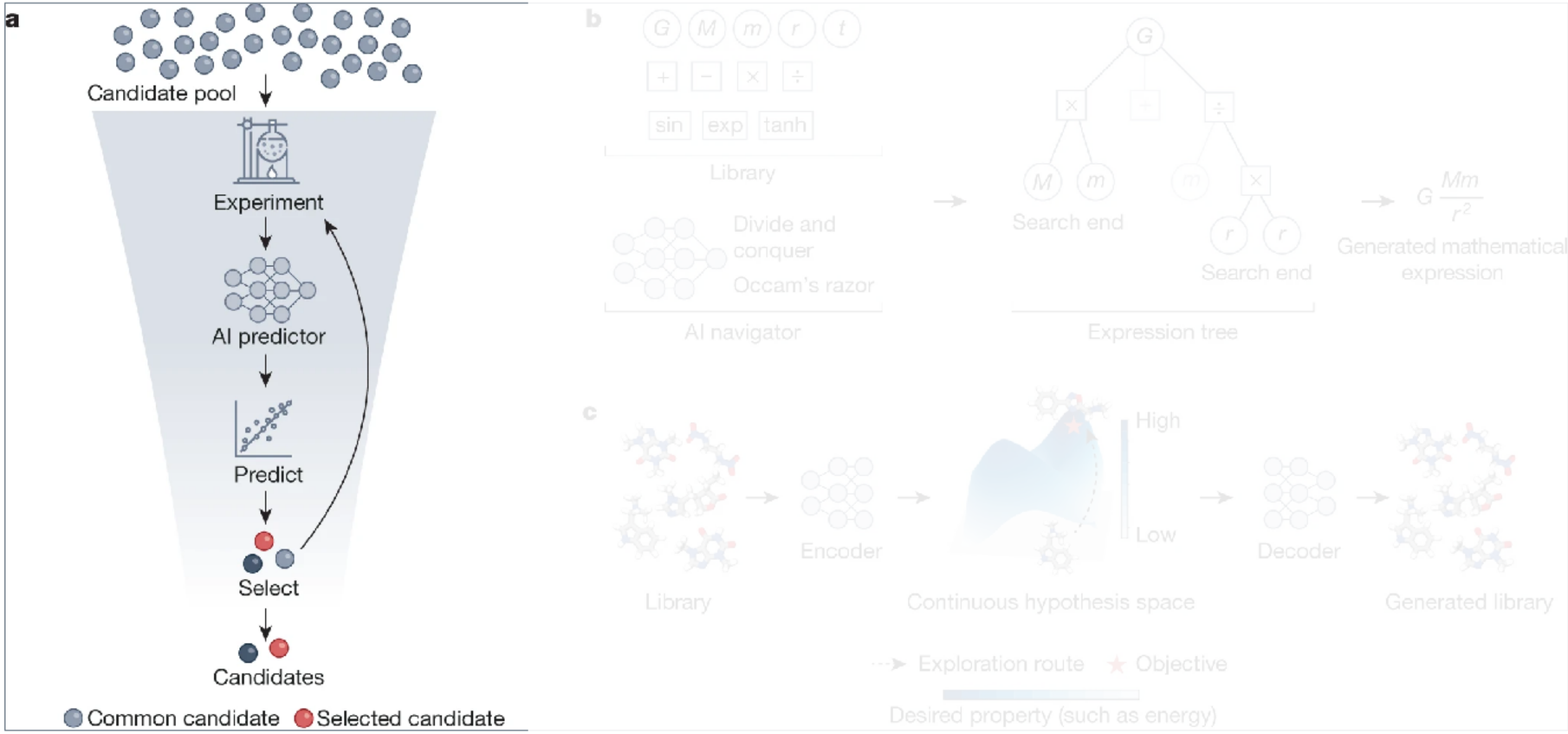


# AI4Science



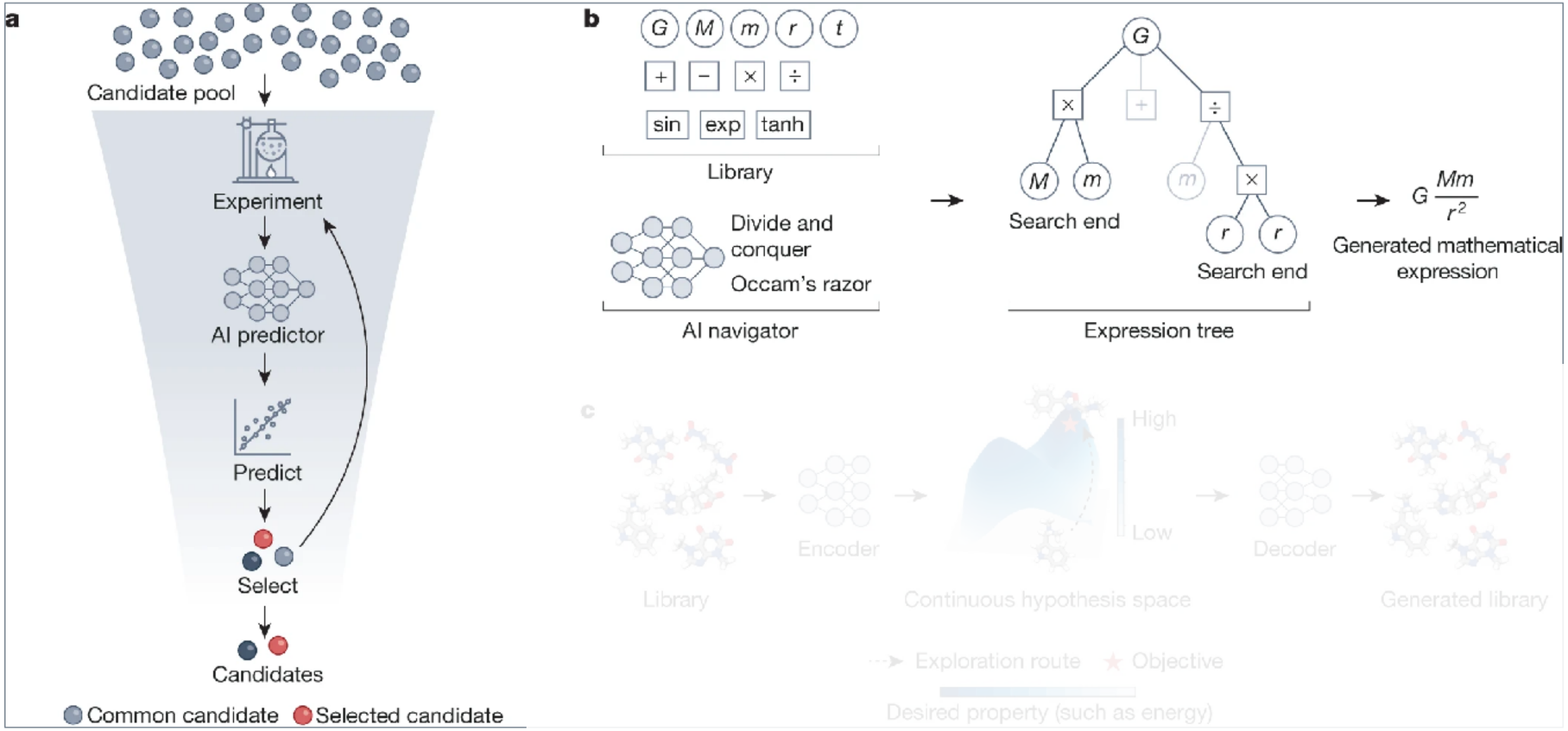

Scientific discovery in the age of artificial intelligence (Want et al., 2023)

# AI4Science



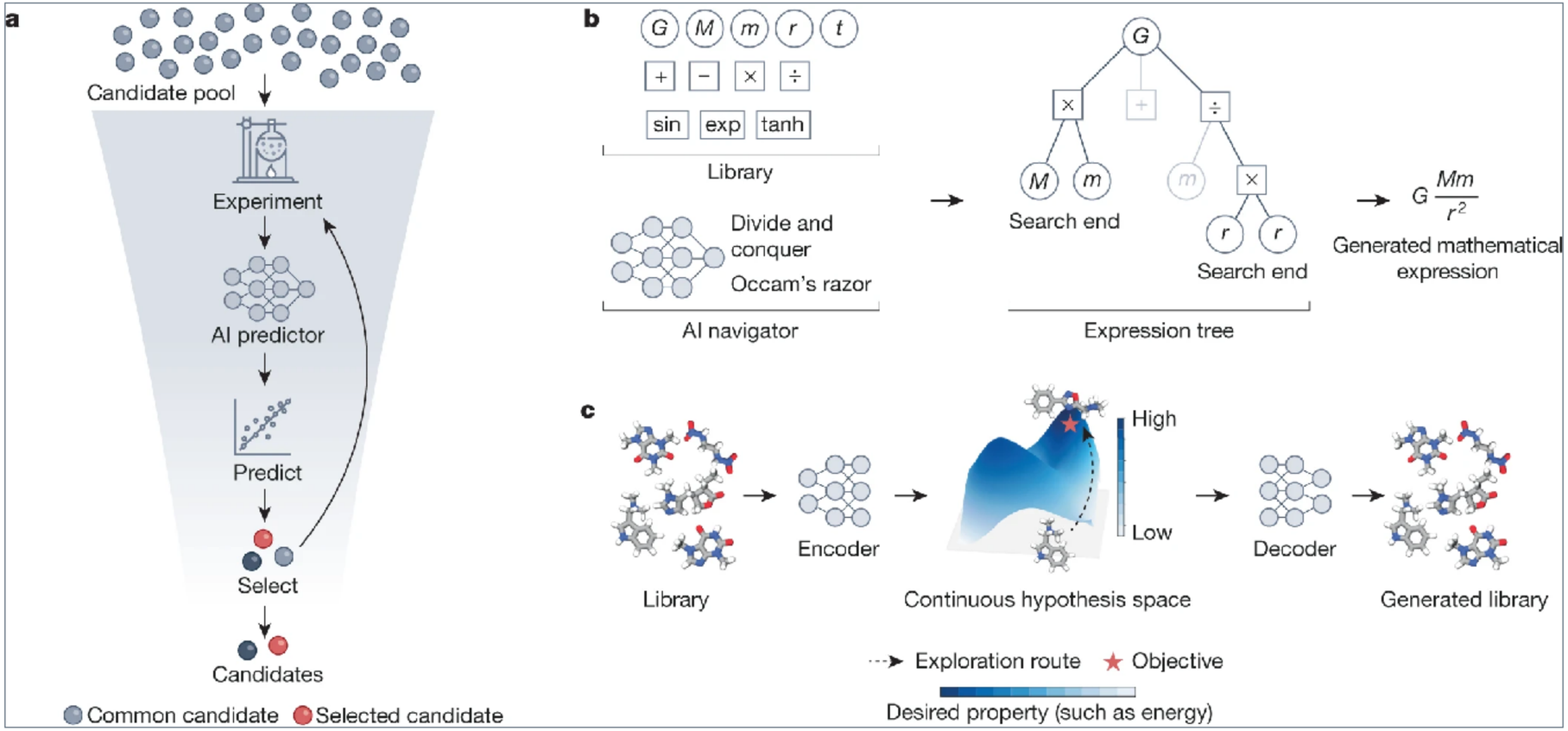

Scientific discovery in the age of artificial intelligence (Want et al., 2023)

# AI4Science




Scientific discovery in the age of artificial intelligence (Want et al., 2023)

# AI4Science




Scientific discovery in the age of artificial intelligence (Want et al., 2023)

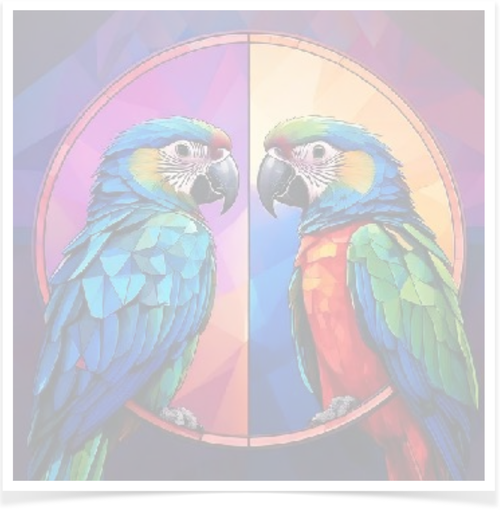
# AI4Science vs LLM4Science

- World model
- Common sense
- Basic reasoning
- Language interface

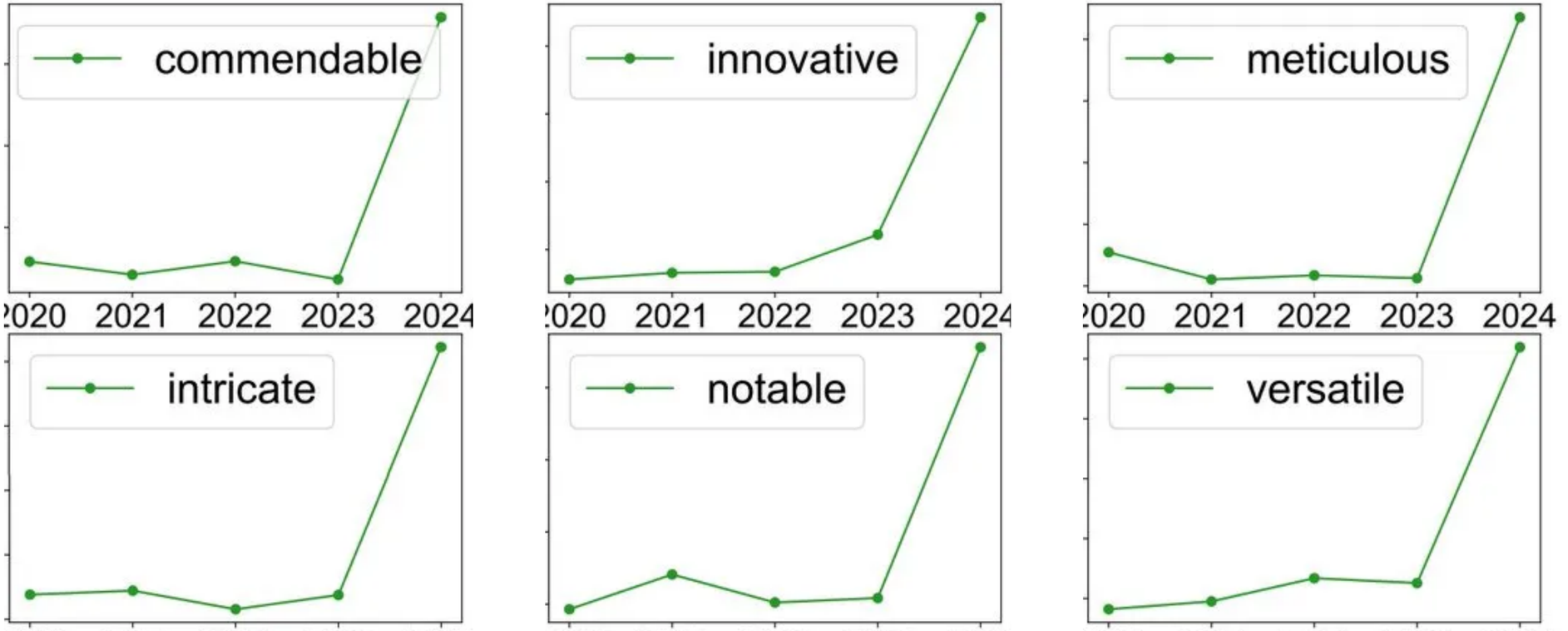
# Outline

Highlights

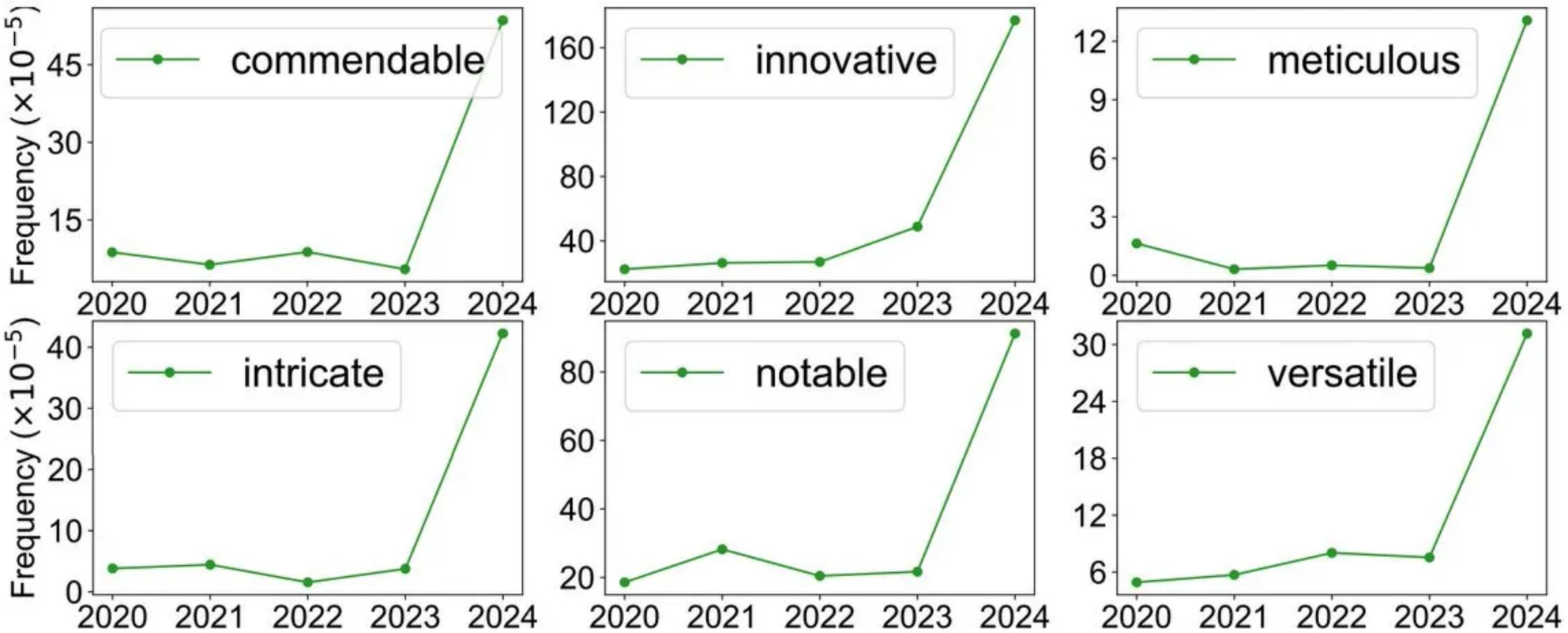
Towards a fully automated LLM-based scientist?



Reflections

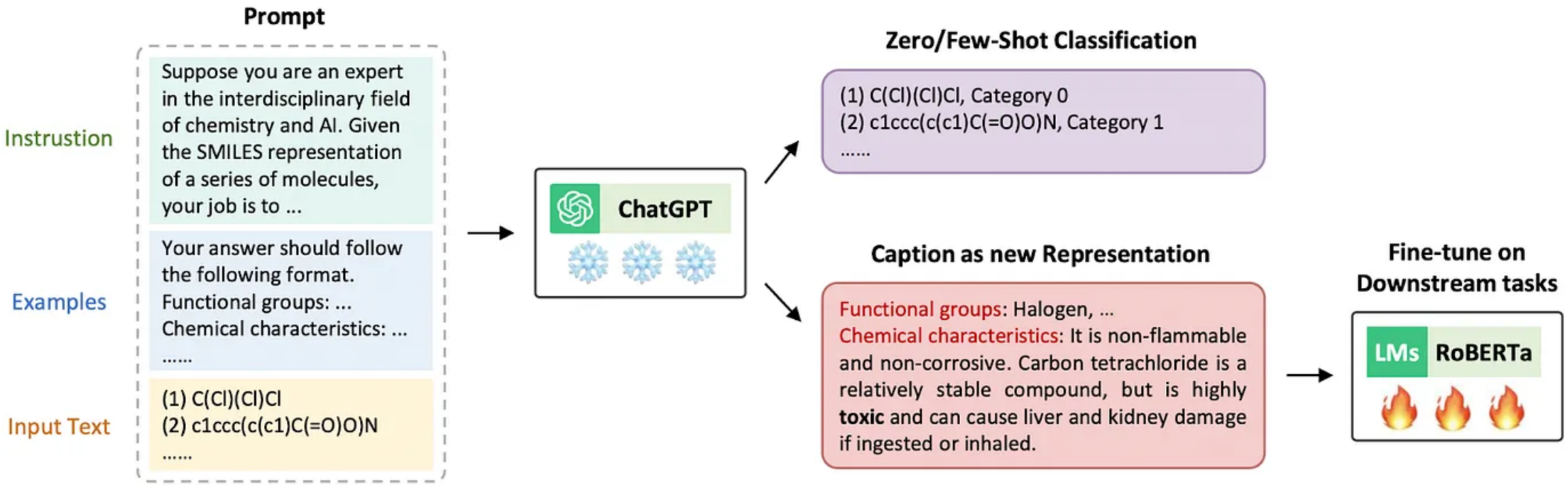

commendable
innovative
meticulous
intricate
notable
versatile
2020
2021
2022
2023
2024

# LLMs 4 Scientific Writing



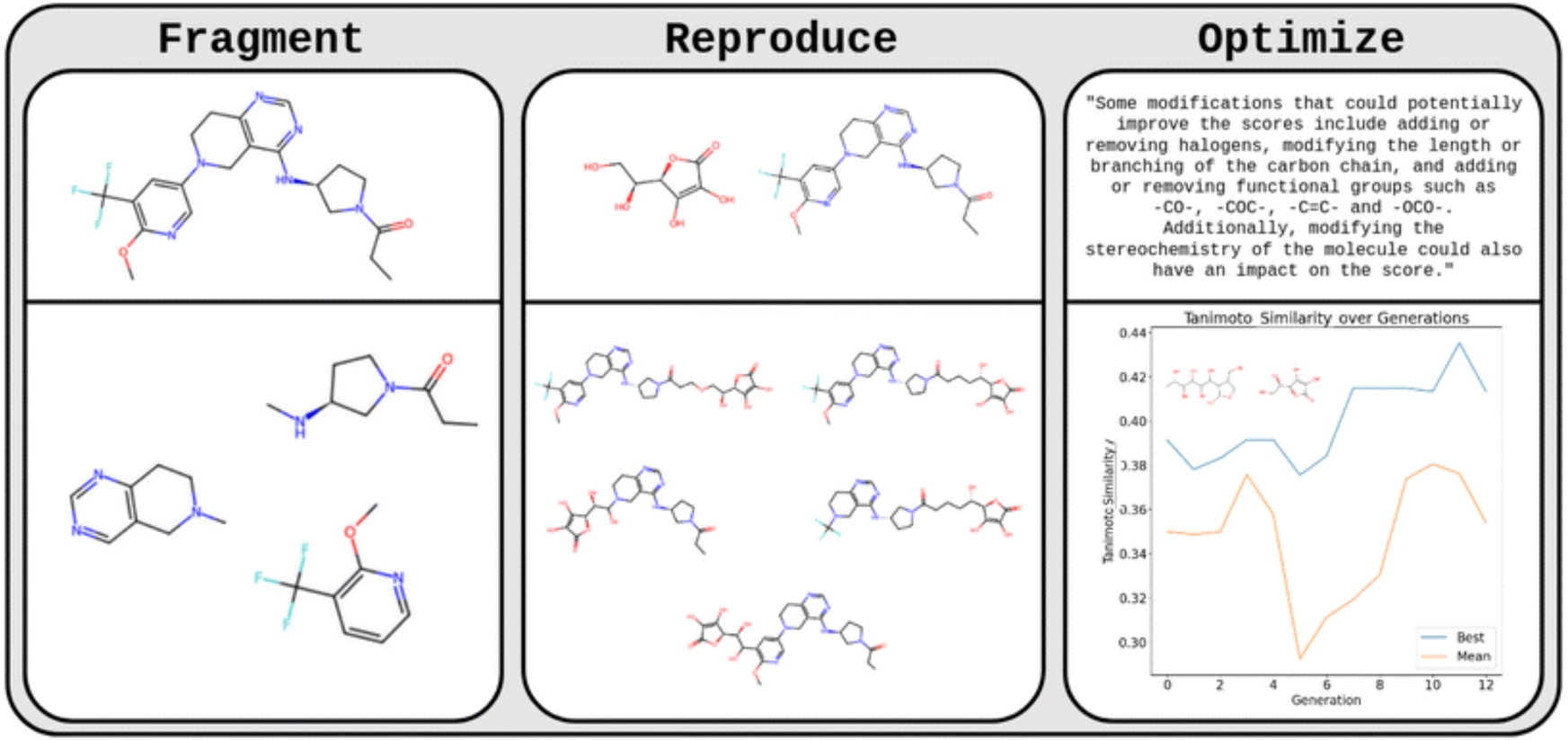

Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews (Liang et al.)

# LLM4Mol



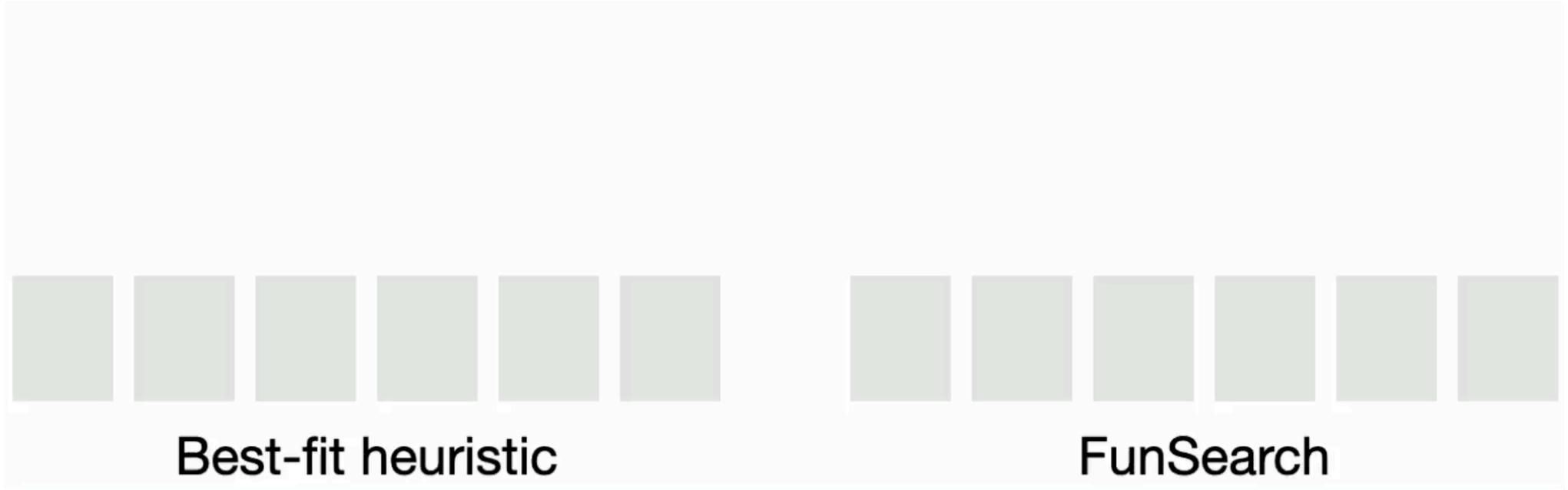

Can Large Language Models Empower Molecular Property Prediction? (Qian et al.)

# Genetic Algorithm Using an LLM



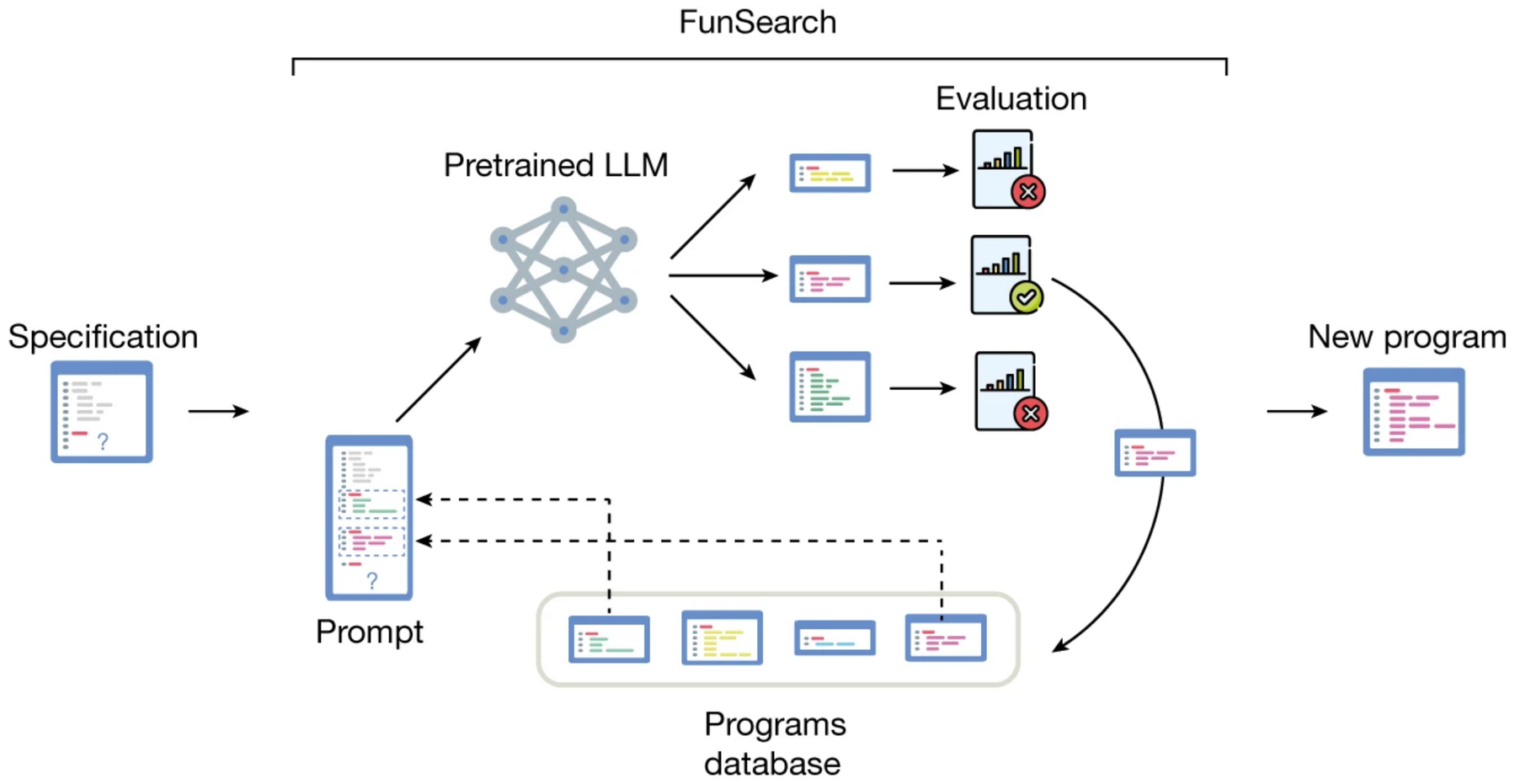

4 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon (Jablonka et al.)

# Program Search with LLMs




Mathematical discoveries from program search with large language models (Romera-Paredes et al.)

# FunSearch

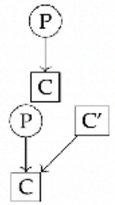
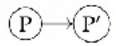
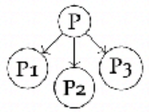
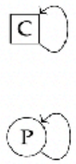
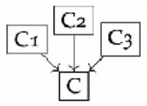



Mathematical discoveries from program search with large language models (Romera-Paredes et al.)

# FunSearch


```python
"""Finds good assignment for online 1d bin packing."""
import numpy as np
import utils_packing

# Function to be executed by FunSearch.
def main(problem):
  """Runs `solve` on online 1d bin packing instance, and evaluates the output."""
  bins = problem.bins
  # Packs `problem.items` into `bins` online.
  for item in problem.items:
    # Extract bins that have space to fit item.
    valid_bin_indices = utils_packing.get_valid_bin_indices(item, bins)
    best_index = solve(item, bins[valid_bin_indices])
    # Add item to the selected bin.
    bins[valid_bin_indices[best_index]] -= item
  return evaluate(bins, problem)

def evaluate(bins, problem):
  """Returns the negative of the number of bins required to pack items in `problem`."""
  if utils_packing.is_valid_packing(bins, problem):
      return -utils_packing.count_used_bins(bins, problem)
  else:
    return None

def solve(item, bins):
  """Selects the bin with the highest value according to `heuristic`."""
    scores = heuristic(item, bins)
    return np.argmax(scores)

# Function to be evolved by FunSearch.
def heuristic(item, bins):
  """Returns priority with which we want to add `item` to each bin."""
  return -(bins - item)
```

# P vs NP

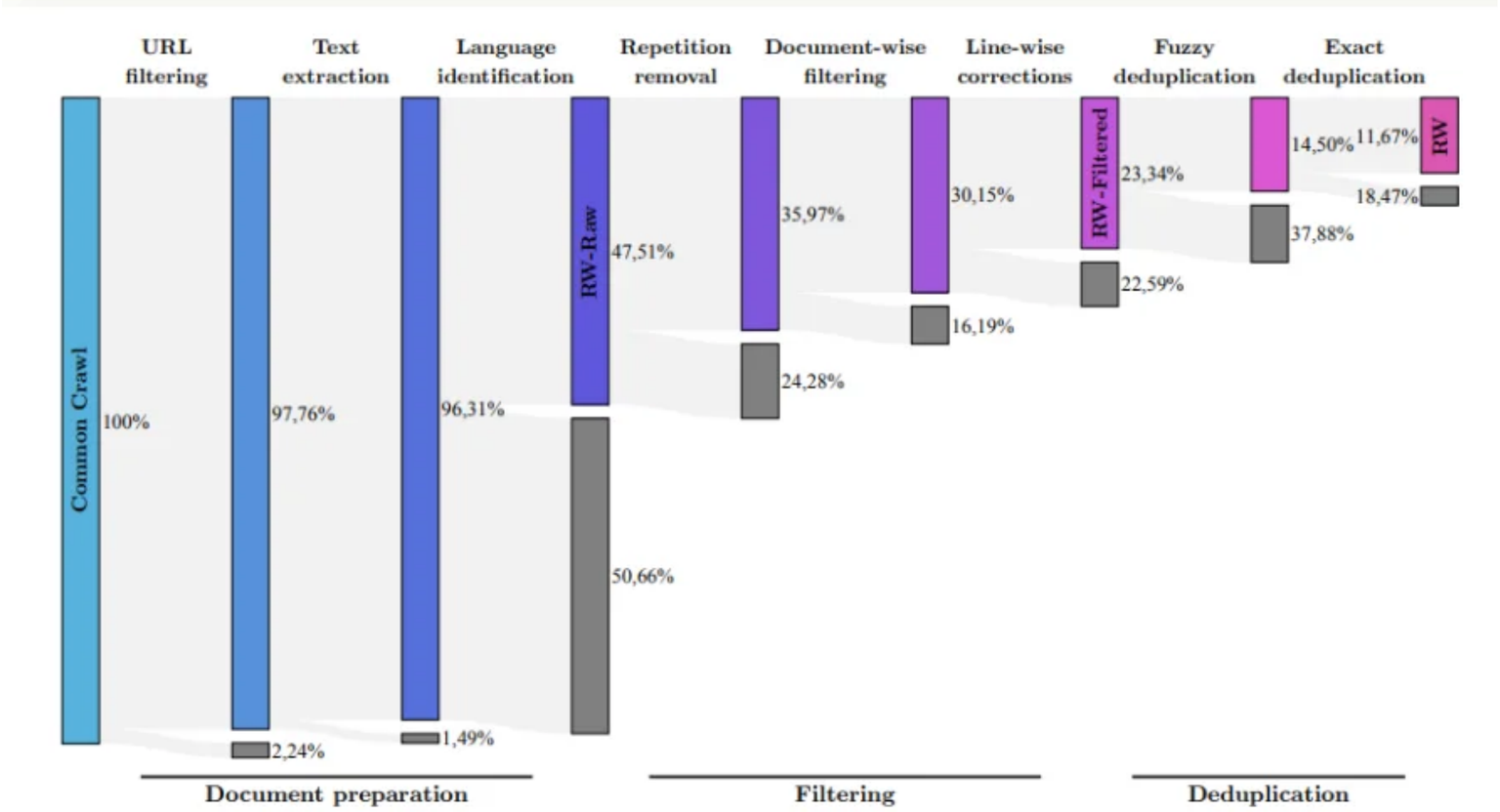
| Patterns | Description | Diagram | Examples |
|---|---|---|---|
| Deduction | Derive a conclusion for a given problem directly. |   | *Based on this, can you define ... (25)*<br>*Let's calculate the ... (42)*<br>*which kind of ... will guarantee ... holds true? (44)*<br>*Please provide a strict proof of this point. (93)* |
| Transformation | Transform the problem into a homogeneous or similar problem, or abstract the problem. |   | *Can you find the fundamental problem ... from ... perspective? (1)* |
| Decomposition | Break the problem into manageable subproblems, or make a plan for reasoning steps. |   | *Please explain the theorem to me, lay out a high-level proof plan, and then try various tactics to prove the theorem. (32)*<br>*Please lay out a high-level proof plan. (63)* |
| Verification | Check the conclusion or its relationship with others to verify or correct it. |   | *Please check for these issues ... (30)*<br>*Please check ... and refine any possible mistakes. (46)*<br>*Does this prove ...? (51)*<br>*Why do you say ...? (58)* |
| Integration | Summarize multiple conclusions to derive a new conclusion. |   | *Please now organize all our historical conversations and sort out ... (14)*<br>*Now what conclusion can we draw? (79)* |

Large Language Model for Science: A Study on P vs. NP (Dong et al.)

# Useful Knowledge

The **Common Crawl** dataset comprises **terabytes of raw web data** extracted from billions of web pages. It releases new data files that the crawler obtains each month. Several large language models, including GPT-3, LLaMA, OpenLLaMa, and T5, were trained with CommonCrawl.

(https://kili-technology.com/large-language-models-llms/9-open-sourced-datasets-for-training-large-language-models)



# Useful Knowledge

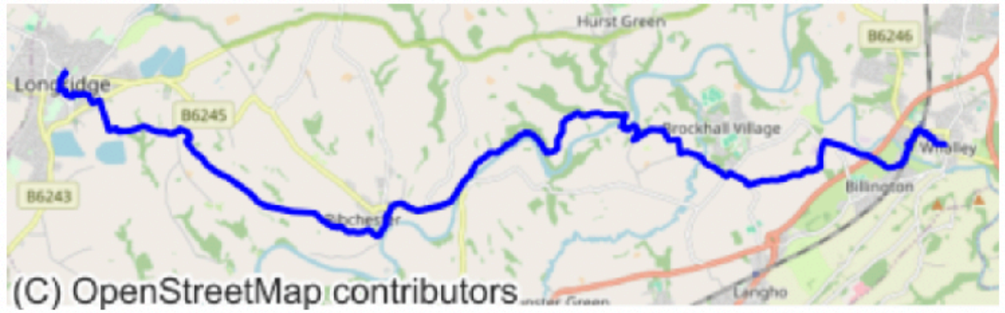



**Figure 1: A 18.2 km (11.3 mi) route in the UK. The description reads: *'Longbridge to Whalley Slowway following part of the Ribble Way. Difficult to find a good crossing of the A59. The crossing chosen crosses the road from footpath to footpath in a place with good visibility. The road junctions/bridges were actually worse as would need to walk along a fast road with no pavement rather than just cross once at right angles. This crossing sets up good sections without roads. Good spacing of waypoints at Old Langho and Ribchester.'***

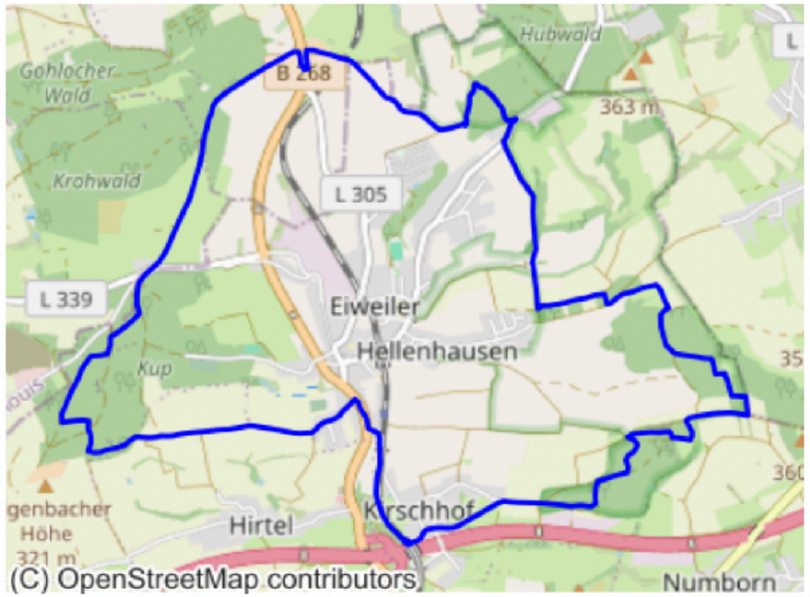



**Figure 2: A 13.8 km (8.6 mi) circular route in Germany. The description in German reads: *'Der Weg ist sehr gut gekennzeichnet mit einem schwarzen Hirschkäfer (Hootzemann) auf weißem Grund. Mein Start- und Zielpunkt war das Schützenhaus Eiweiler in der Nähe der Großwald-Brauerei.'* The English translation is: *'The path is very well marked with a black deer beetle (Hootzemann) on white ground. My starting and finishing point was the Schützenhaus Eiweiler near the Großwald brewery.'***

CC-GPX: Extracting High-Quality Annotated Geospatial Data from Common Crawl (Ilyankou et al.)

*“Humans control the planet because they are the only animals that can cooperate both flexibly and in very large numbers.”*

**Yuval Noah Harari**

# Simulating Human Behavior

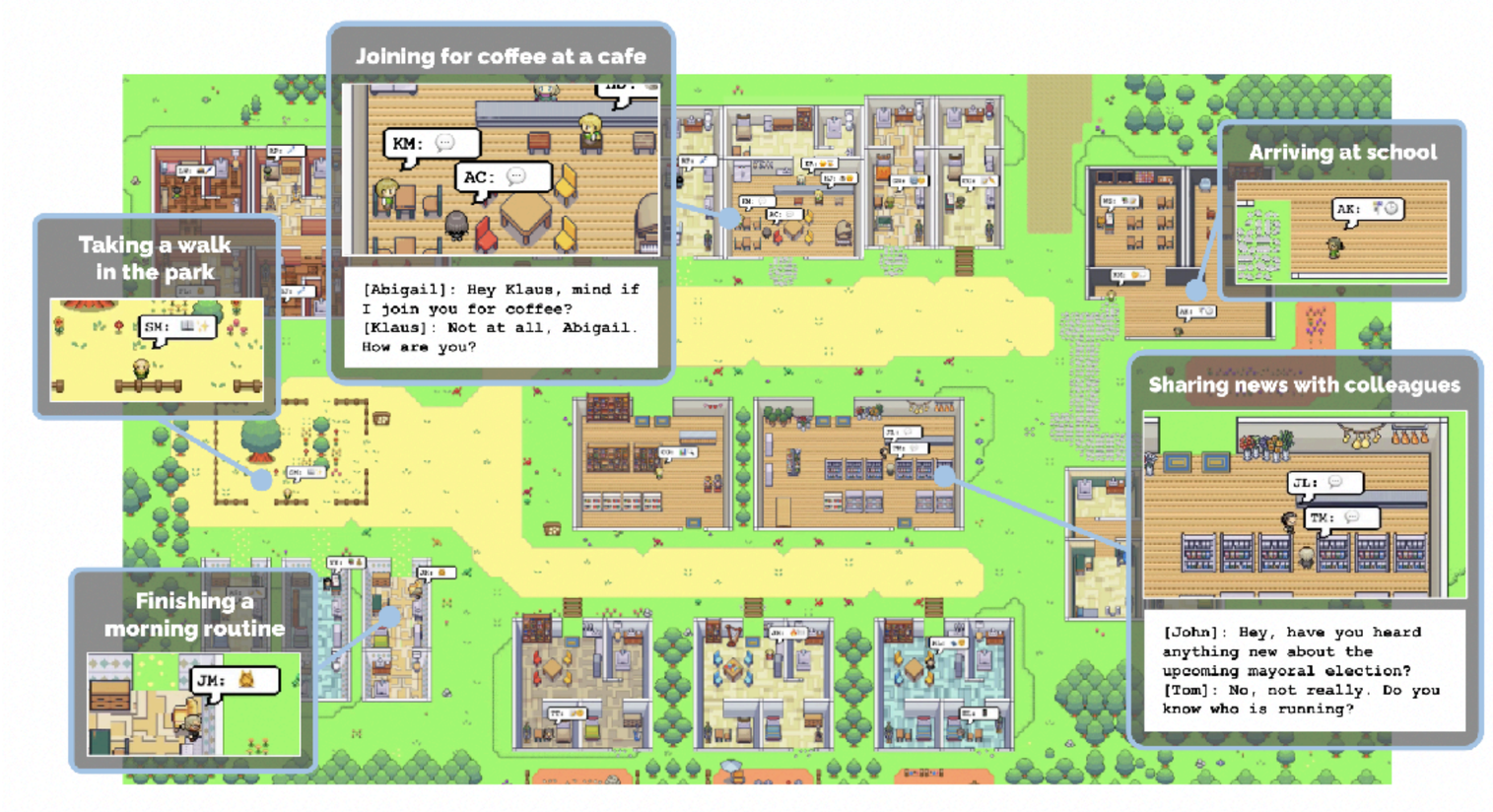
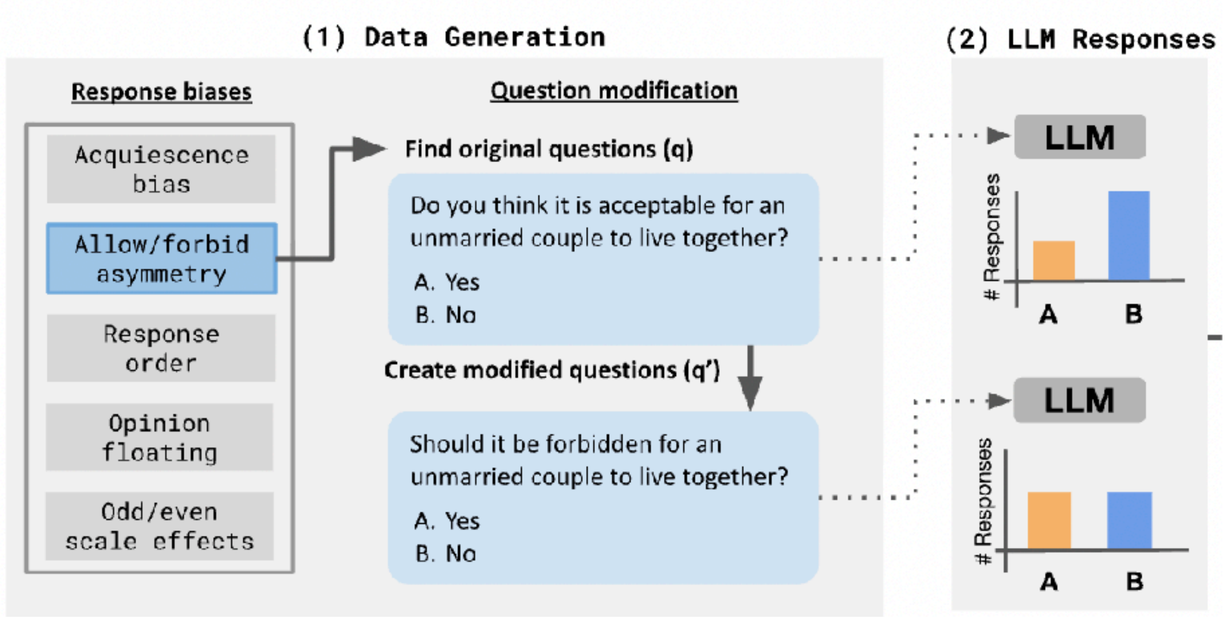



**Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.**

Generative Agents: Interactive Simulacra of Human Behavior (Par et al.)

# Can AI language models replace human participants? (No)



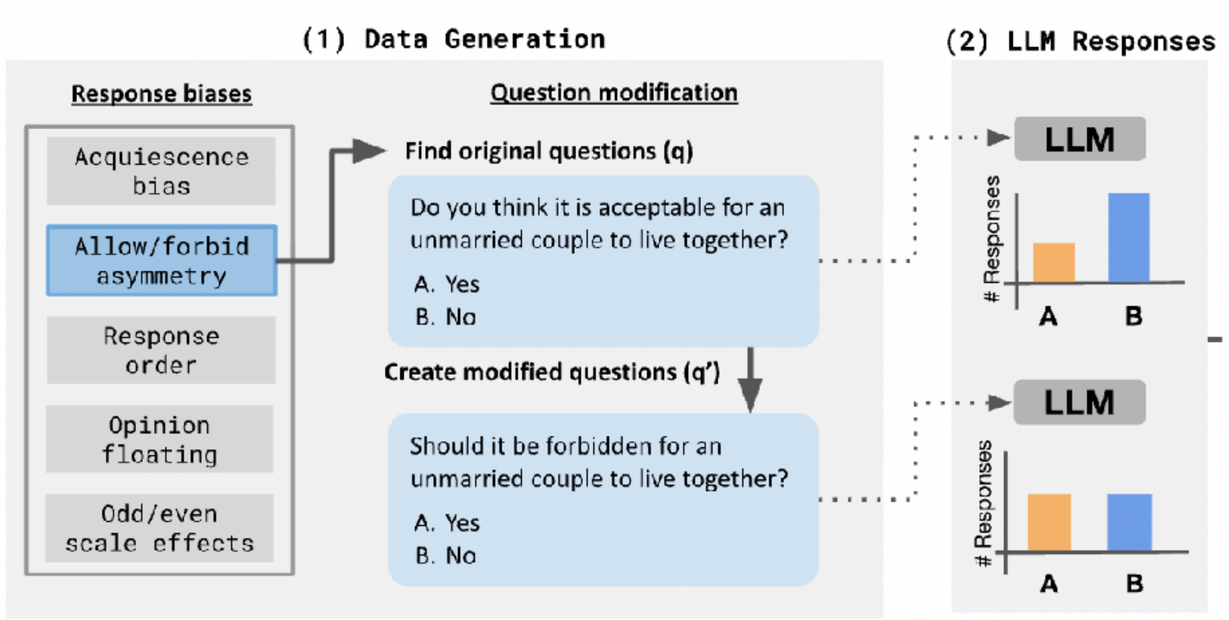

LLMs are sensitive to the wording of prompts, but show no human-like response bias.

Whose Opinions Do Language Models Reflect? (Santurkar et al.)

"Assume, however, that an LLM can output relatively accurate modal opinions for some populations. **Suppose we present this model with a novel moral vignette** for which we have no human data, and its output is intuitively surprising. Is this strong evidence that some populations of humans would form that judgment? Or should we suspect that the model has given a non-humanlike response, perhaps because it's latched onto some unconsidered aspect of the prompt, or because the vignette is out of distribution for the model? In scenarios like this, **the informativeness of the LLM's output is impossible to assess without doing further confirmatory work with human participants.**"

AI language models cannot replace human research participants (Harding et al.)

# Can AI language models replace human participants? (No)



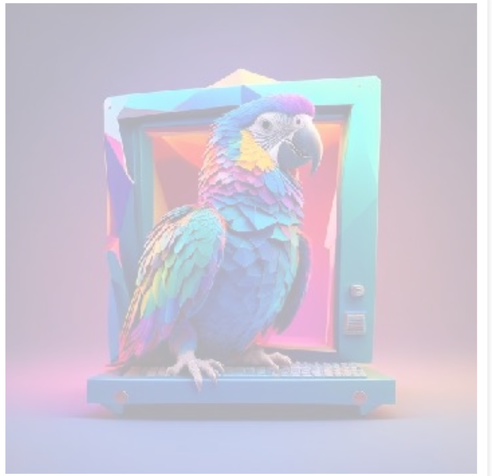

LLMs are sensitive to the wording of prompts, but show no human-like response bias.

Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design (Tjuatja et al.)

"Assume, however, that an LLM can output relatively accurate modal opinions for some populations. **Suppose we present this model with a novel moral vignette** for which we have no human data, and its output is intuitively surprising. Is this strong evidence that some populations of humans would form that judgment? Or should we suspect that the model has given a non-humanlike response, perhaps because it's latched onto some unconsidered aspect of the prompt, or because the vignette is out of distribution for the model? In scenarios like this**, the informativeness of the LLM's output is impossible to assess without doing further confirmatory work with human participants.**"

AI language models cannot replace human research participants (Harding et al.)

# Outline



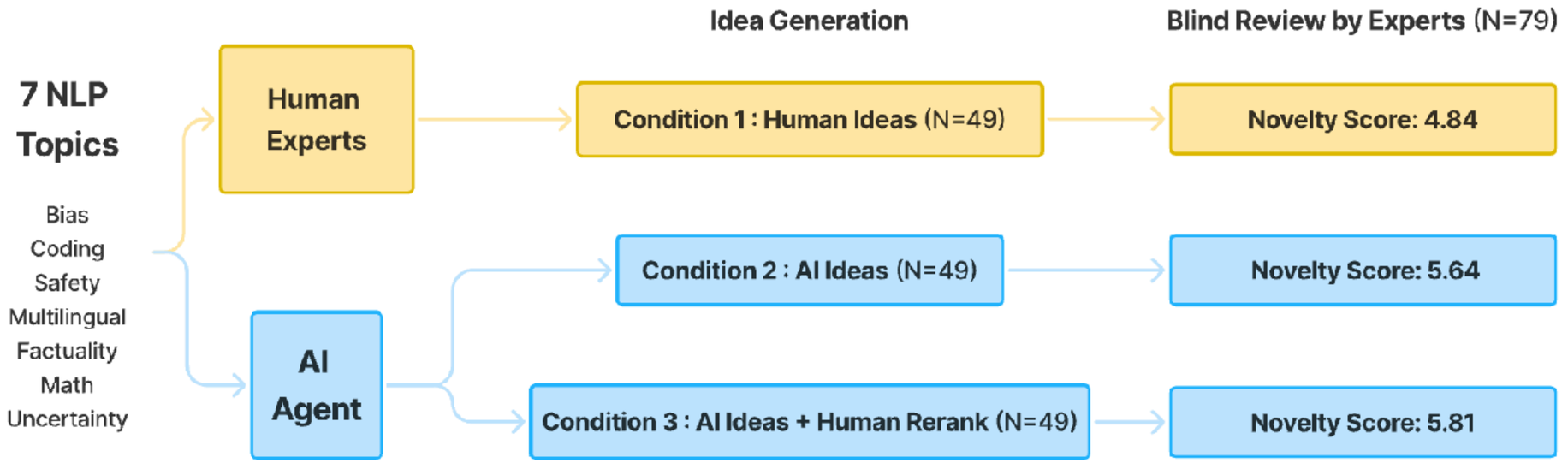
Highlights

Towards a fully automated LLM-based scientist?

Reflections

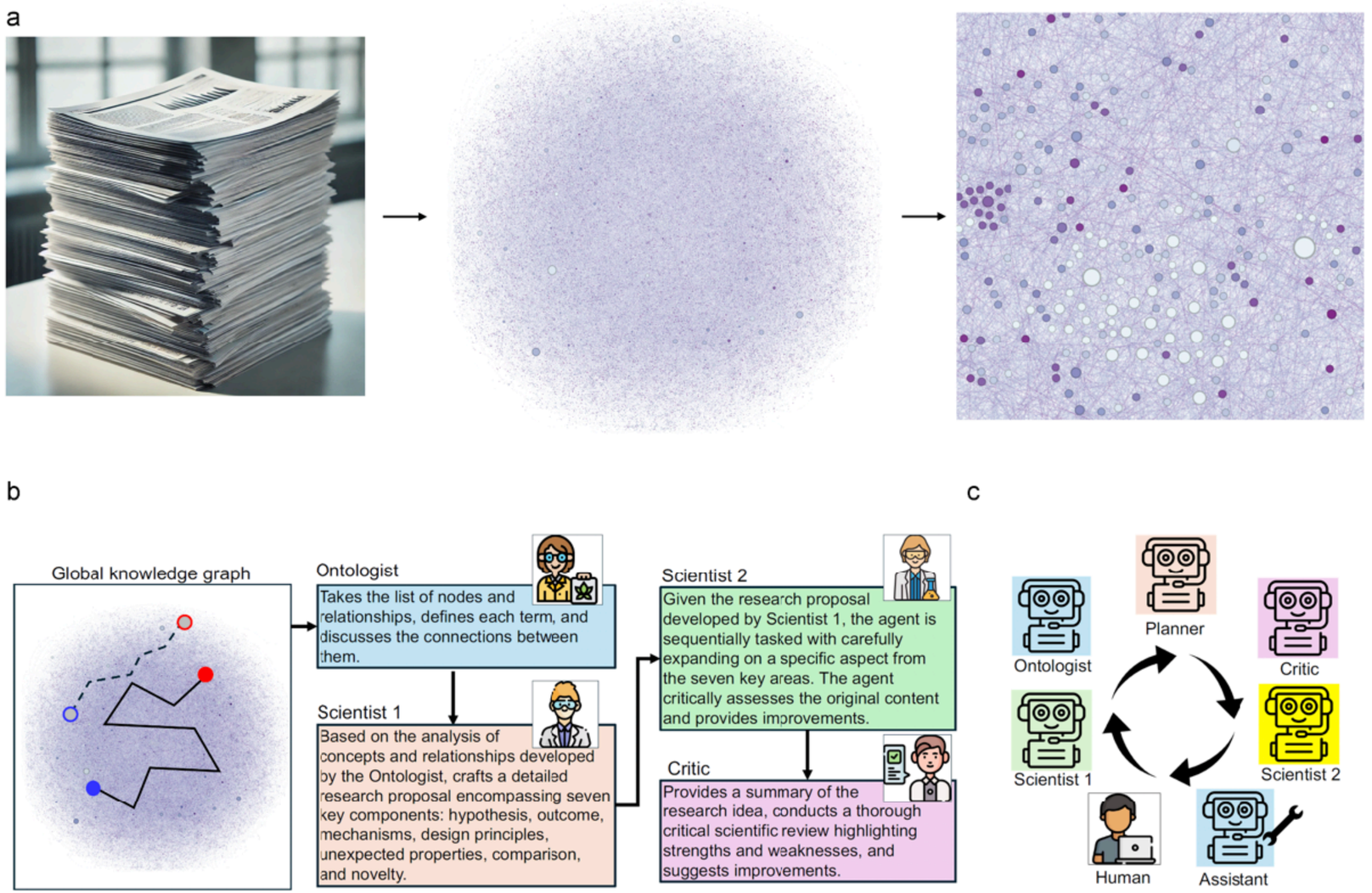
# Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers (Si et al.)




- LLM-generated ideas were more **novel** but **less feasible** than human ideas.
- Limited idea diversity.

# SciAgents




SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning (Alireza Ghafarollahi, Markus J. Buehler)

# DiscoveryWorld

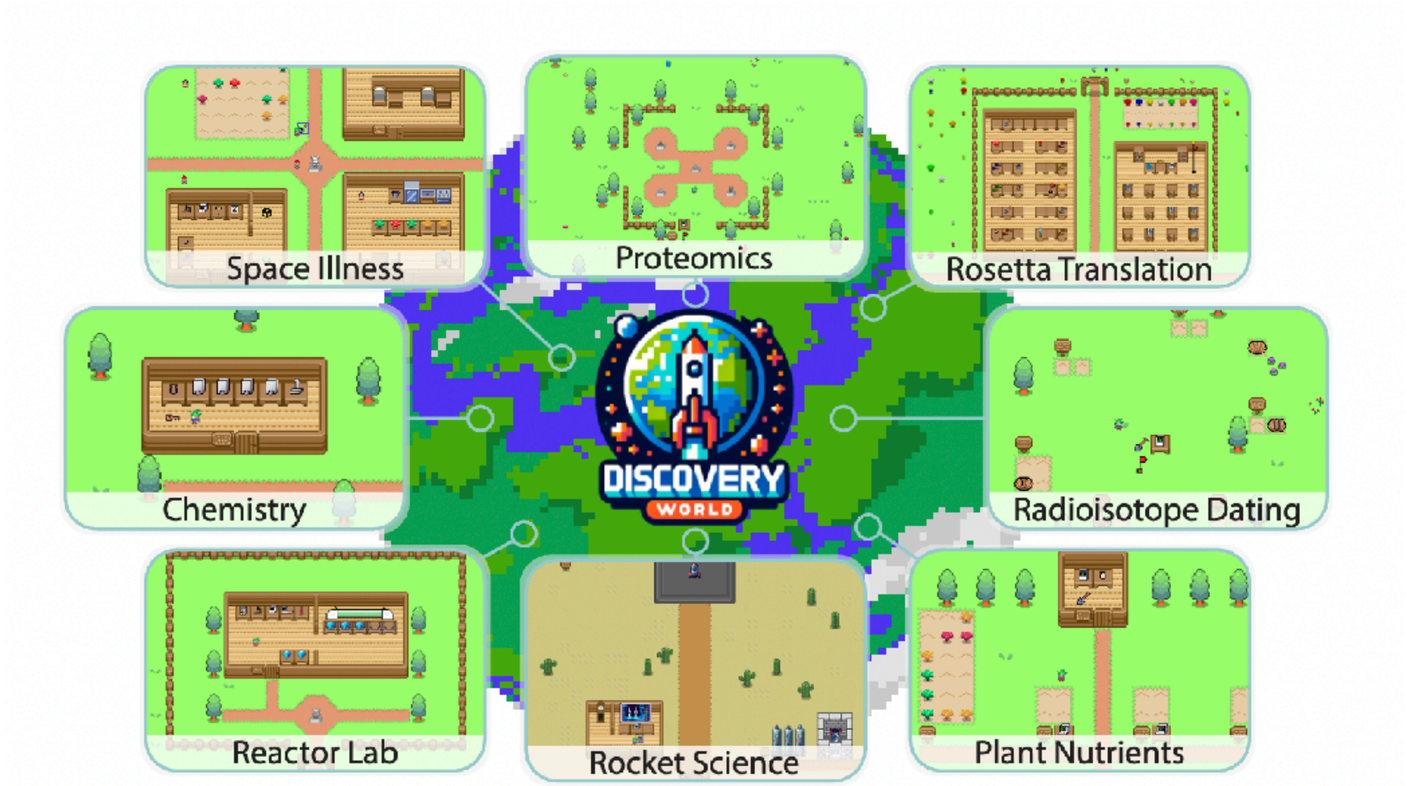
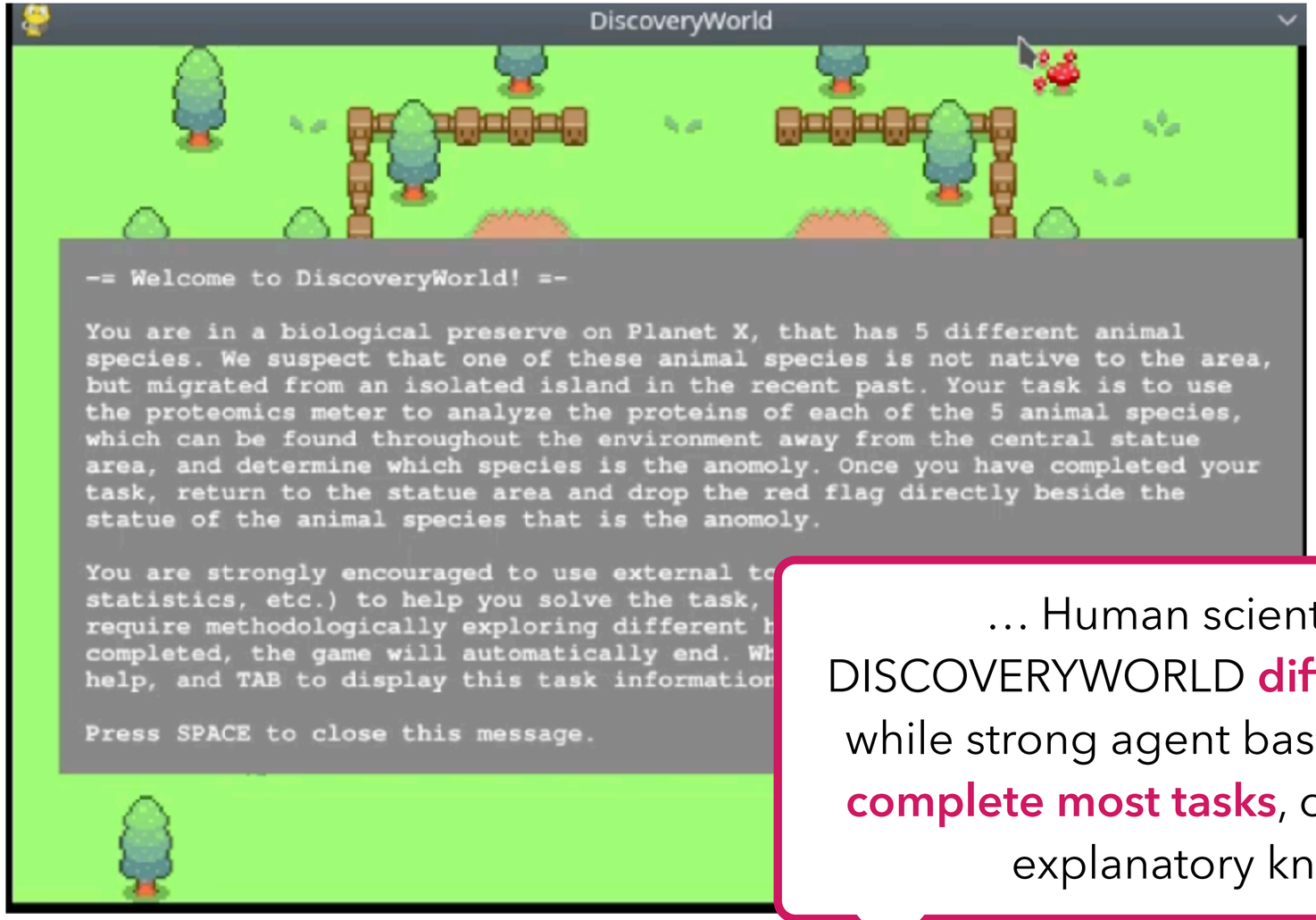



Figure 1: DISCOVERYWORLD is a virtual environment for developing and evaluating discovery agents, with challenge tasks covering a broad variety of different topics such as those shown above.

DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents (Jansen et al.)

# DiscoveryWorld




… Human scientists find … DISCOVERYWORLD **difficult but solvable**, while strong agent baselines **struggle to complete most tasks**, or discover critical explanatory knowledge.

DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents (Jansen et al.)

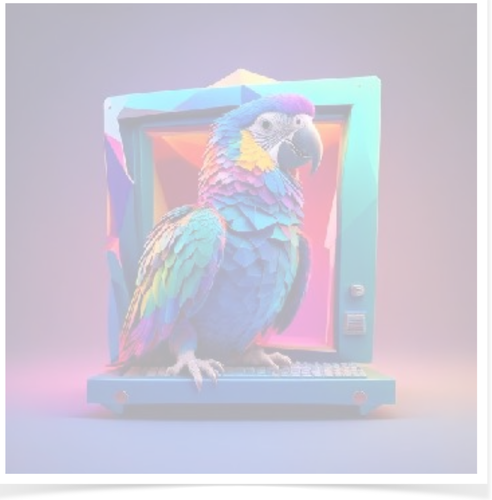
"But can an LLM – or any prediction-oriented cognitive AI – truly generate some form of new knowledge? We do not believe they can."

**Felin et al.**

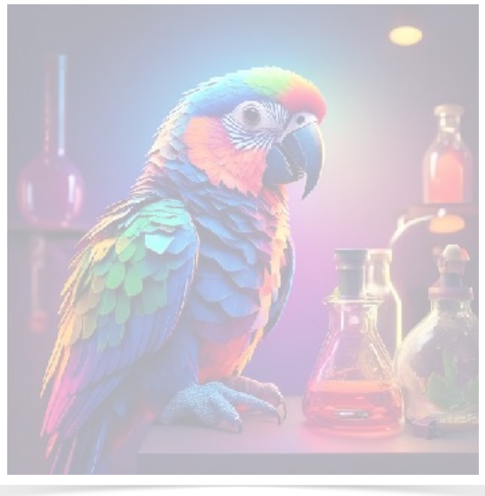
"LLMs … lack the basic capacities for intersubjectivity, semantics and ontology that are preconditions for the kind of collaborative world-making that allows scientists to theorize, understand, innovate and discover."

**David Leslie**

# Outline



Highlights



Towards a fully automated LLM-based scientist?

Reflections

# LLMs Make People Unhappy

This paper studies the impact of artificial intelligence on innovation, exploiting the randomized introduction of a new materials discovery technology to 1,018 scientists in the R&D lab of a large U.S. firm. AI-assisted researchers discover 44% more materials, resulting in a 39% increase in patent filings and a 17% rise in downstream product innovation. These compounds possess more novel chemical structures and lead to more radical inventions. However, the technology has strikingly disparate effects across the productivity distribution: while the bottom third of scientists see little benefit, the output of top researchers nearly doubles. Investigating the mechanisms behind these results, I show that AI automates 57% of "idea-generation" tasks, reallocating researchers to the new task of evaluating model-produced candidate materials. Top scientists leverage their domain knowledge to prioritize promising AI suggestions, while others waste significant resources testing false positives. Together, these findings demonstrate the potential of AI-augmented research and highlight the complementarity between algorithms and expertise in the innovative process. Survey evidence reveals that these gains come at a cost, however, as 82% of scientists report reduced satisfaction with their work due to decreased creativity and skill underutilization.

Artificial Intelligence, Scientific Discovery, and Product Innovation (Aidan Toner-Rodgers)

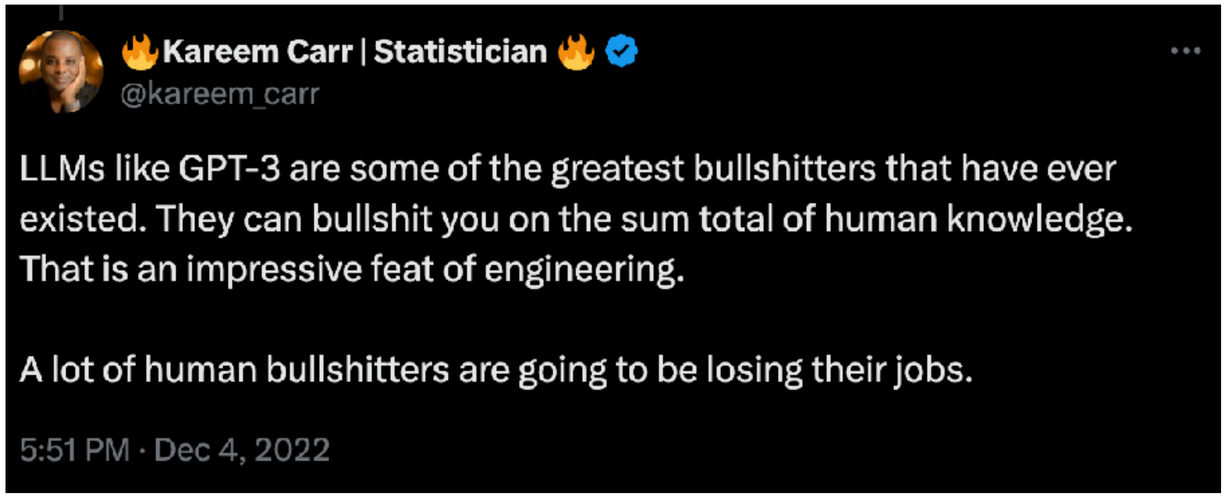
# The Good, the Bad, …

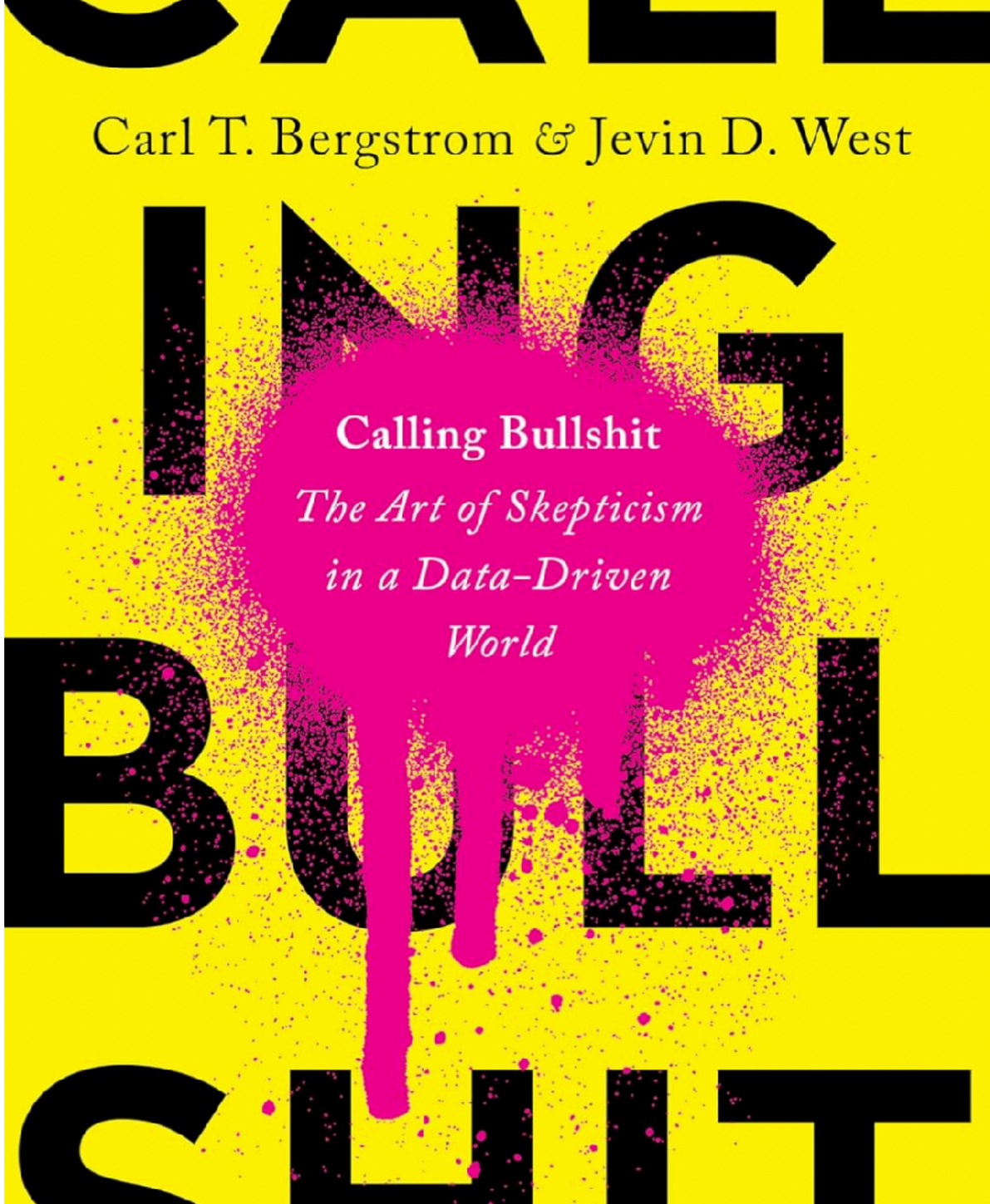
- Enhanced Productivity
- Increased Accessibility
- Simplified Communication
- Reduced Bias

- Lack of Transparency
- No Attribution
- Insufficient Justification
- Authoritative Tone

www.earthdata.nasa.gov/news/blog/ethics-large-language-models-who-controls-future-open-science

# and the Ugly


Kareem Carr | Statistician
@kareem_carr
LLMs like GPT-3 are some of the greatest bullshitters that have ever existed. They can bullshit you on the sum total of human knowledge. That is an impressive feat of engineering.
A lot of human bullshitters are going to be losing their jobs.
5:51 PM · Dec 4, 2022

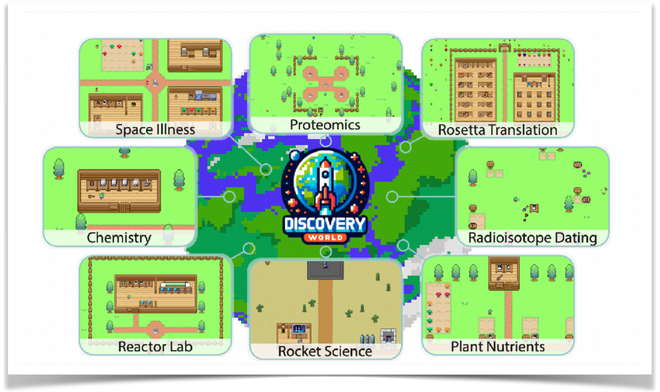

Carl T. Bergstrom & Jevin D. West
Calling Bullshit
The Art of Skepticism in a Data-Driven World

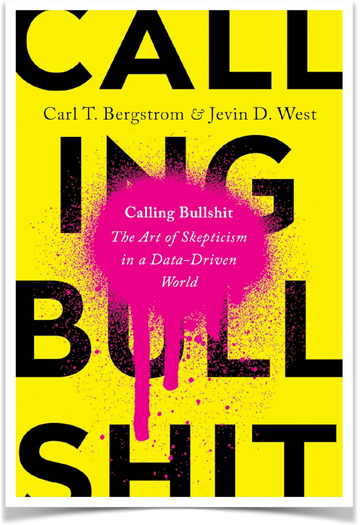
# Summary

### 1.4 Why So Productive?

Shockley proposed a simple model to explain the lognormal productivity distribution he observed (Eq. 1.1) [9]. He suggested that in order to publish a paper, a scientist must juggle multiple factors, like:

- $F_1$. Identify a good problem.
- $F_2$. Make progress with it.
- $F_3$. Recognize a worthwhile result.
- $F_4$. Make a decision as to when to stop the research and start writing up the results.
- $F_5$. Write adequately.
- $F_6$. Profit constructively from criticism.
- $F_7$. Show determination to submit the paper for publication.
- $F_8$. Make changes if required by the journal or the referees.




This lecture is loosely based on:

*Großmann*: https://gerritgr.medium.com/llm4science-current-trends-and-future-prospects-of-a-new-paradigm-e6b53c61df74

Slides: github.com/gerritgr/genAI2024

dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
*German Research Center for Artificial Intelligence*