Position: Data-driven Discovery with Large Generative Models

Bodhisattwa Prasad Majumder [*1] **Harshit Surana** [*2] **Dhruv Agarwal** [*3] **Sanchaita Hazra** [4]
Ashish Sabharwal [1] **Peter Clark** [1]

Abstract

With the accumulation of data at an unprecedented rate, its potential to fuel scientific discovery is growing exponentially. This position paper urges the Machine Learning (ML) community to exploit the capabilities of large generative models (LGMs) to develop automated systems for end-to-end *data-driven discovery*—a paradigm encompassing the search and verification of hypotheses purely from a set of provided datasets, without the need for additional data collection or physical experiments. We first outline several desiderata for an ideal data-driven discovery system. Then, through DATAVOYAGER, a proof-of-concept utilizing GPT-4, we demonstrate how LGMs fulfill several of these desiderata—a feat previously unattainable—while also highlighting important limitations in the current system that open up opportunities for novel ML research. We contend that achieving accurate, reliable, and robust end-to-end discovery systems solely through the current capabilities of LGMs is challenging. We instead advocate for fail-proof tool integration, along with active user moderation through feedback mechanisms, to foster data-driven scientific discoveries with efficiency and reproducibility.

1. Introduction

The deluge of data collected in the digital age by advanced scientific instruments, sensors, and computational techniques has marked a transformative change in the process and pace of scientific discovery (Anderson, 2008; Ramakrishnan & Grama, 1999; Jumper et al., 2021). This acceleration, however, paints a paradoxical scenario—while rapid development indicates the advancement of knowledge, it simultaneously poses significant challenges for scientists to absorb new findings, navigate interconnections, formulate novel hypotheses, and arrive at meaningful conclusions (Bianchini et al., 2022). To facilitate future scientific progress, it is, therefore, imperative to develop automated systems that are capable of continuous ingestion, creative generation, and analytical reasoning at a massive scale.

Developing an end-to-end discovery system is challenging. Previous works have either severely lacked the requisite computational power (Langley, 1981; Langley et al., 1984; 1983), developed domain-specific bespoke methodologies (e.g., AlphaFold; Jumper et al. (2021)), or involved substantial human intervention (e.g., wet lab experiments) thus not qualifying as autonomous end-to-end (CoScientist; Boiko et al. (2023)). In this position paper, we argue that **a focus on data-driven discovery using large generative models (LGMs)** addresses each of these prior shortcomings and presents a practical first step towards the goal of an end-to-end system for automating the scientific process. Following Newell & Simon (1976), we define this paradigm as a heuristic search framework that aims to describe a given set of observations by uncovering the laws that govern its data-generating process.

For example, consider the flow described in Figure 1. Given a dataset of socio-economic variables collected from a set of respondents, a user might formulate a hypothesis about the relationship between the BMI of a subset of the respondents and their financial behavior (variables present in the dataset; *top-left*). A data-driven discovery system should be able to automatically generate a verification plan and execute multiple steps of statistical tests (e.g., OLS, GLM) over the provided data to confirm or reject the hypothesis (*bottom-left*). Alternatively, a user might only provide a high-level research question, such as specifying the domains of interest (i.e., finance and health; *top-middle*). In this scenario, a discovery system must first identify the relevant variables and then search the space of plausible hypotheses to generate and verify interesting questions conditioned on the provided data and existing world knowledge (*bottom-middle*). Finally, users may have diverse information-seeking needs necessitating the ability to provide feedback to the system, such as in using a particular statistical methodology for certain types of data during automatic verification (*top-right*). An

[*]Equal contribution [1]Allen Institute for AI [2]OpenLocus [3]University of Massachusetts Amherst [4]University of Utah. Correspondence to: Bodhisattwa Prasad Majumder <bodhisattwam@allenai.org>, Harshit Surana <harshit@openlocus.ai>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

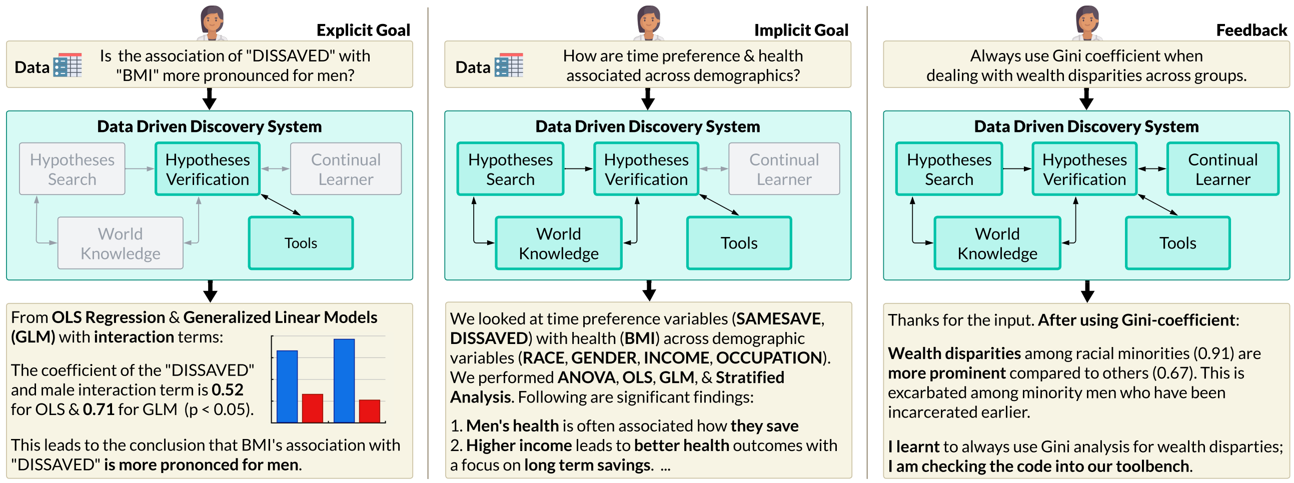

Figure 1. **A blueprint flow** demonstrating ideal workflows for data-driven discovery. **Left:** User asks an explicit question around a particular line of inquiry or hypothesis. **Middle:** The user can also ask a broad and partially-defined high-level question, where the system must figure out the appropriate datasets, data transformations, variables, a list of possible hypotheses, and their verification. In this example, the system maps time preference and health outcomes to exact variables, runs the analysis across appropriate demographic cuts, and then shares the significant findings for further exploration and verification. **Right:** The user can provide follow-up feedback at any time and the continual learner will learn from it while providing updated experiments and results.

automated discovery system must accommodate and persist such feedback in order to recover from mistakes and accurately handle future queries (*bottom-right*).

While our ultimate goal encompasses the full spectrum of scientific inquiry, we focus first on end-to-end discovery from observational or experimental data for two reasons: (1) an abundance of large-scale datasets that would benefit highly from automated discovery; and (2) the practicality of automated verification enabled by data without the need for additional data collection[1].

We identify two main challenges to automating data-driven discovery—(1) *hypothesis search*: the effective consumption of provided data and existing knowledge to devise novel hypotheses, and (2) *hypothesis verification*: the evaluation of the generated hypotheses for rapid iteration and continual discovery. A successful solution must further be able to generate and follow complex plans, execute diverse analytical tests, and parse through the abundant heterogeneity in real-world data. With the unprecedented success of LGMs operating on multiple modalities such as language (Achiam et al., 2023; Touvron et al., 2023), code (Liu et al., 2023b; Li et al., 2022), and images (Achiam et al., 2023; Liu et al., 2023a), we argue that it is now practical to build such a solution that can effectively tackle both challenges.

Hypothesis Search. The scientific process typically begins with the construction of a proposed hypothesis based on prior knowledge and exploratory observations regarding some phenomenon of interest. For example, discovering new insights from publicly available National Longitudinal Surveys[2] will require prioritizing unexplored hypotheses over already verified results.

Foremost, we may ask whether the search should be driven by an extrinsic goal—a user-defined objective, a high-level research question, or a set of variables of interest. This setting might involve using algorithms that guide the search process using objective-gradients (Weitzman, 1978) that identify variables and models that directly, or *greedily*, optimize the extrinsic goal. We argue that LGMs, with their massive, web-scale pre-training, possess both the necessary priors and the ability to handle heterogeneity, to help guide such a goal-driven search for relevant hypotheses.

It can also be argued that goal-driven approaches may not yield desired outcomes, particularly when dealing with open-ended questions, where the search is often susceptible to capture by local optima (Whitley, 1991; Bengio et al., 2009). Drivers for search might then be intrinsic metrics (Oudeyer et al., 2007), such as diversity (Eysenbach et al., 2019; Agarwal et al., 2023; Trinh et al., 2024), interestingness (or curiosity) (Pathak et al., 2017; Zhang et al., 2023), or information gain (Hennig & Schuler, 2012; Houthooft et al., 2016), that do not optimize for a user-defined extrinsic goal but instead encourage open-ended creativity and, eventually, serendipitous discovery (Foster & Ford, 2003; Taleb, 2007; Stanley et al., 2017). Here, too, LGMs present a solu-

[1] In contrast to hypothesis verification in the physical sciences, which often require wet lab experiments and where erroneous automation may lead to false discoveries (Leeman et al., 2024).

[2] `https://www.bls.gov/nls/`

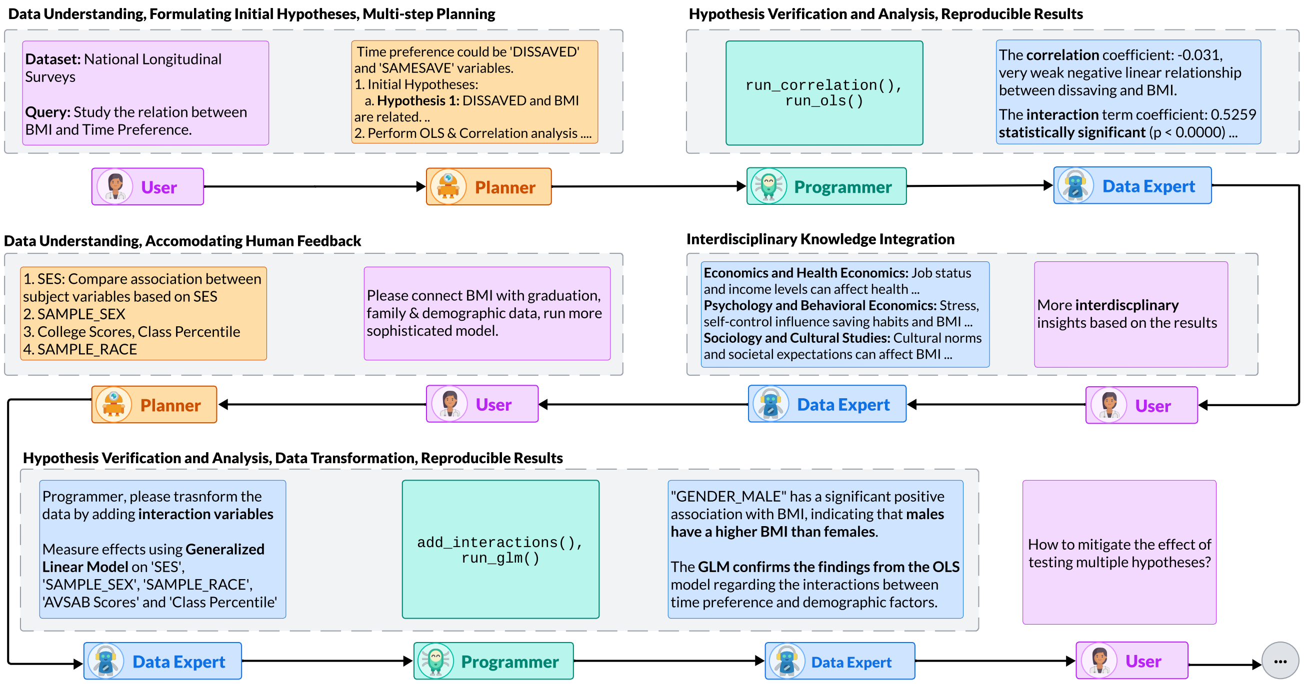

Figure 2. **An example workflow** of DataVoyager. Starting from a user-provided dataset and a high-level query, it navigates through cycles of hypothesis generation, validation, and analysis to uncover complex insights. See **all examples in Appendix** for full understanding.

tion, for instance, in estimating the novelty or likelihood of hypotheses in the search space.

Hypothesis Verification. With a set of plausible hypotheses identified, it is next required to subject each claim through detailed inspection, often via a series of empirical evaluations and statistical tests, to determine veracity, which is highly tractable and could be fail-proof in data-driven discovery. This might involve selecting which analyses or statistical tests to run, transforming raw data into a format admissible for each test, handling missing or erroneous data, generating code to execute the tests, and finally analyzing the test results. Given the surge of recent advancements in language modeling capabilities, including instruction-following (Wei et al., 2022), tool use (Schick et al., 2023), program synthesis (Wang et al., 2023a; Agarwal et al., 2023), planning (Majumder et al., 2023), and orchestration (Hou et al., 2023), we argue that LGM agents present a promising solution for automating hypothesis evaluation.

The availability of these capabilities, however, must not be seen as a panacea. (1) LGMs often hallucinate, leading to incorrect insights that may not be grounded in the data. (2) LGMs have limited or no "System-2" reasoning (Kahneman, 2011; LeCun, 2022; Kambhampati et al., 2024), thus necessitating additional scaffolding in order to utilize them for long-horizon tasks. (3) LGMs demonstrate subpar performance in the long tail, thus making their successful application in interfacing with external and domain-specific tools a major challenge to overcome. (4) Finally, LGMs are notoriously challenging to align and steer based on human feedback (Wolf et al., 2023), a crucial component for reliable and useful scientific discovery.

We envision a blueprint of a data-driven discovery system in Figure 1 that allows researchers to ingest datasets, search and verify hypotheses using fail-proof tools, and consult literature to surface novel insights. Our survey in Figure 3 indicates the lack of systems capable of automated and robust data-driven discovery, with existing systems partially covering desired functionalities. To tackle this, we argue:

1. Automated data-driven discovery warrants research attention owing to the abundance of (public or private) data and its tractable challenges (hypothesis search and verification) as opposed to discoveries requiring laborious data collection or physical experiments.
2. LGMs present an incredible potential to realize several properties of an ideal data-driven discovery system, such as knowledge-driven hypothesis search or tool usage to verify hypotheses—creating new avenues for ongoing efforts in the ML community on code generation, planning, and program synthesis.
3. LGMs are *not* all we need. Interfacing with fail-proof tools and inference-time functions, catering to domains and long tail with user moderation, is required to have an accurate, reliable, and robust data-driven discovery system capable of advancing scientific progress with speed and reproducibility.

2. DATAVOYAGER: A Proof of Concept

As a proof of concept, we borrow a well-studied role-based multi-agent architecture (Liu et al., 2023c; Zhou et al., 2023) powered by GPT-4 (Achiam et al., 2023), a state-of-the-art language model, to build DATAVOYAGER—a system that can semantically understand a dataset, programmatically explore verifiable hypotheses using the available data, run basic statistical tests (e.g., correlation and regression analyses) by invoking pre-defined functions or generating code snippets, and finally analyze the output with detailed analyses. DATAVOYAGERis meant to represent a *baseline system* that utilizes existing functionalities of GPT-4, such as function calling, code generation, and language generation.

We envision any data-driven discovery system to be capable of operating in either of the following two settings. (1) *Fully-autonomous*: using only the dataset and its metadata as the input. In this case, the system should consider the full hypothesis space for search and verification. (2) *User-guided*: combining the dataset with a (natural language) query stating a high-level objective to narrow down the hypothesis search space, akin to goal-directed agents (Majumder et al., 2023). DATAVOYAGER can operate in both settings.

The core components of our system consist of specialized agents that are designed to manage different aspects of the data-driven discovery process as well as structured functions or programs that help analyze the data in specific ways via function calling. We employ the AutoGen framework[3] that allows agents to communicate in arbitrary order dependent on the context. Following is a brief description of all agents used in DATAVOYAGER (more in Figure 4):

- **Planner:** Interprets the user query and generates a comprehensive, structured plan to achieve it or, in the autonomous setting, generates an additional dataset exploration plan. The plan is then decomposed into executable sub-tasks and delegated to the relevant agents.
- **Programmer:** Performs data transformations, filtering, and specialized coding for domain-specific analyses according to the generated plan. It can also call structured, pre-defined functions with relevant arguments to make execution fail-proof.[4]
- **Data Expert:** Interprets the results generated by the programmer, extracting insights, connecting interdisciplinary knowledge, and formulating conclusions.
- **Critic:** Evaluates the analyses and provides constructive feedback on analytical methods and execution.
- **User Proxy:** Facilitates on-demand human feedback. A user can steer the discovery process towards an objective, rectify errors, and prevent off-course explorations.

[3] https://microsoft.github.io/autogen/

[4] We develop several functions (e.g., statistical analysis tools based on datatype, python shell execution tools) for robustness.

3. Towards Data-driven Discovery Systems

In this section, we first outline a set of desired **functionalities** for a data-driven discovery system. Using these functionalities, and armed with our baseline system DATAVOYAGER along with evidence from the literature, we demonstrate extensive support towards our positions 2 and 3. Functionalities such as data understanding, hypothesis generation, multi-step planning, and interdisciplinary knowledge integration provide evidence that a system (DATAVOYAGER) powered by a state-of-the-art LGM shows promise for ideal data-driven discovery, an ability not previously achievable before the wide adoption of LGMs. On the other hand, functionalities such as data transformation, scale, hypothesis verification, accommodating human feedback, and p-hacking proof confirm that LGMs alone are insufficient. Integrating robust tools to execute at scale and user-centric interventions is crucial for a tractable data-driven discovery system.

3.1. Comprehensive Data Understanding

Data Understanding. Understanding data forms the core of data-driven discovery and involves processing variables semantically as well as programmatically (Ristoski & Paulheim, 2016). The former involves understanding how the data was collected or acquired, grounded in the domain knowledge, as well as the semantic relationship between the variables present in the data. The latter involves understanding the datatype of each variable and the values they can take. Progress in database query generation (Sun et al., 2023), commonsense reasoning on symbolic spaces (Qiu et al., 2023), and unsupervised KGQA (Agarwal et al., 2023) indicate reason for optimism for the use of LGMs for data understanding.

For example, Smith et al. (2005) explored the link between time preference and BMI from the National Longitudinal Surveys using several variables indicating the saving behavior of the respondents. To replicate this from scratch, a discovery system must understand the difference between the variable measuring if respondents withdrew more money from savings than they put in (DISSAVED) and the variable indicating if they have no savings or unchanged savings from the previous year (SAMESAVE)[5]. Here, DATAVOYAGER's LGM-based planner correctly identifies variables related to time preference:

> To examine the effects of time preference on individuals, we need to understand the variables in the dataset that relate to time preference. In the provided dataset, the variables **`DISSAVED`** and **`SAMESAVE`** seem to be related to time preference (...) *Full example:* Figure 6

[5] Time preference reflects how individuals value present over future benefits. A lower time preference can lead to higher savings, better food consumption, and thus a healthier BMI in the future.

While it works for this example, a comprehensive data understanding is still challenging due to the complexity of various datasets with numerous types and complex metadata. We, therefore, ask: *can a system achieve a comprehensive understanding of domains and variations in diverse datasets in a domain-agnostic manner as compared to domain-specific systems, such as CoScientist (Boiko et al., 2023)?*

Data Transformation. Different datasets have unique characteristics, requiring custom transformations and filtering operations (Kang et al., 2017). Moreover, even within the same dataset, different hypotheses may demand different transformations for accurate verification and testing. Without this capability, the potential to conduct a wide range of statistical tests for hypothesis verification would be compromised (Bailis et al., 2017). A simple example of data transformation would be the ability to convert a categorical variable into a one-hot encoding. Further, the following is an example showing DATAVOYAGER's LGM-based programmer performing data transformation in order to derive interaction terms between variables:

> Let's start by adding interaction terms to examine the potential link between time preference and BMI across different demographic groups (. . .) *Full example*: Figure 7

The challenge lies in accommodating the abundant diversity of hypotheses and datasets, each requiring highly customized transformations (Bowers & Ludäscher, 2004). The ability of LGMs to generate code for such domain-specific data (Sharma et al., 2023) hints towards a generalized solution; however, the difficulty in debugging generated code (Vaithilingam et al., 2022) demands a call to action for building better code generation models.

Scale. Modern scientific exploration often involves large amounts of data, a complex analytics workflow, and a large hypothesis space (Elliott et al., 2016). It is important, thus, for a useful autonomous discovery system to be able to sift through such large datasets efficiently while maintaining the state of its several processes and tracking previously conducted analyses. Without this ability to scale and handle complex workflows, several hypotheses would remain unexplored, and valuable insights left undiscovered.

For longitudinal studies, where it is important to understand how variables evolve over time (Weiss & Ware, 1996), scalability is particularly crucial in order to handle data over extended time periods. Furthermore, in very large-scale data scenarios, such as the Cancer Moonshot project[6] and the Cancer Genomics Cloud (Lau et al., 2017), the discovery system must be able to analyze petabytes of data in complex workflows, all while maintaining a state of the possible hypotheses and variable combinations as well as the explorations conducted thus far. In such scenarios, LGMs must be able to support long-horizon planning and long-context attention. However, LGMs are yet to show significant progress on both counts (Valmeekam et al., 2022), a limitation of DATAVOYAGER as well, thus highlighting a need for focused research towards these goals.

[6] www.whitehouse.gov/cancermoonshot/

3.2. Hypothesis Generation

Connecting Data and Scientific Literature. The ability to bridge the provided data and existing scientific literature is important in providing an understanding of the hypothesis space grounded by contextual domain knowledge. This ability to learn from known knowledge may further result in various inter-disciplinary perspectives and insights—a phenomenon often called Swanson Linking (Bekhuis, 2006).

For example, to derive novel insights between social background and college graduation (Alexander et al., 1982) from the National Longitudinal Surveys, it is imperative to understand previous research on National Longitudinal Surveys to avoid duplication and incorporate verified knowledge from the literature to improve initial hypotheses.

Linking generated hypotheses to existing knowledge requires accurate retrieval, information extraction, and multistep reasoning (Wang et al., 2023b). Further, combining multiple research articles connects back to the original Swanson Linking problem (Swanson, 1986). While LGMs have recently been shown to perform well in augmenting citations with relevant context based on a user's history (Chang et al., 2023), connecting datasets to scientific literature is an open research problem. By utilizing annotated papers for datasets (Palani et al., 2023), we ask: *can a system learn to combine insights from existing literature and a provided dataset in order to discover novel research gaps?*

Formulating initial hypotheses. Scientists prioritize experiments based on academic intuition, empirical evidence, and existing theories. In data-driven discovery, this approach is akin to selecting hypotheses from a vast combinatorial space of variable interactions, often extensive for exhaustive exploration (Agrawal et al., 2023), to identify dependent and independent variables.

For example, to understand the relationship between *education outcome* and *socioeconomic status*, the system should prioritize investigating how the "rate of completion of BA degree" is influenced by socioeconomic indicators, such as accumulated wealth and parents' education, as a plausible hypothesis (Alexander et al., 1982). This is non-trivial because it not only requires the system to have a semantic understanding of the variable space but also the ability to prioritize hypotheses based on marginal costs and their scientific importance (Agrawal et al., 2023). Here, DATAVOYAGER performs reasonably well on hypothesis generation:

H1: Females are more likely to complete a BA degree compared to males. **H2:** Family size has an impact (...). **H3:** Higher ability scores on the ASVAB test are positively correlated (...) *Full example*: Figure 18

Hypothesis generation can be seen as inductive reasoning (Qiu et al., 2023) using known evidence by connecting them using entailment-like relations (Dalvi et al., 2021). While LGMs show good performance on reasoning benchmarks (Hendrycks et al., 2020), data heterogeneity (e.g., variable names, statistical interactions) and semantics make the reasoning problem harder for LGMs (Lu et al., 2023)—thus, we call for research attention.

3.3. Planning and Orchestrating Research Pathways

Multi-step planning. Data-driven discovery with complex problems and datasets requires a structured approach of breaking down a high-level objective into manageable subtasks, enabling the systematic exploration of the data and hypothesis landscape. This can be considered equivalent to planning (LeCun, 2022). Prioritized hypothesis search with planning involves *states* – the intermediate correlations found from data (sub-hypotheses), and *operators* – the statistical tools and literature to combine verified states (here, sub-hypotheses). Multi-step, iterative planning, thus, comprehensively facilitates the search for scientific discoveries.

Research planning involves incorporating known or novel research pathways, such as the order of analyses or the methods used, and they vary depending on the research goal of the exploration. It can be challenging to choose between a standardized or pre-defined flow as compared to a dynamic plan depending on the realized intermediate states of the planning. Though LGMs as planners are often faulty (Valmeekam et al., 2022), planning within the data hypothesis space presents a fertile ground to systematically benchmark LGMs and improve their abilities.

For example, analyzing the relationship between college education and socio-economic status from National Longitudinal Surveys (Alexander et al., 1982), the system generates the following plan:

I. Understand the data (...) **II.** Generate initial hypotheses (...) **III.** Explore combinations of dependent variables (...) **IV.** Call the "`run_logistic_regression`" function (...) **V.** Repeat step IV for other combinations of dependent variables (...) **VI.** Document the findings (...) **VII.** Seek clarity where required (...). *Full example:* Figure 18

While the ability to decompose abstract plans into executable sub-plans is heavily explored in coding and symbolic reasoning (Khot et al., 2022), DATAVOYAGER presents a strong base case to improve the efficacy of planning by incorporating dynamic strategies that account for search uncertainties.

Exploration vs. exploitation. The debate concerning whether exploration should be goal-oriented or randomized is crucial in making novel discoveries (Agarwal et al., 2023). This applies directly to data-driven discovery, where variable selection by the planner directly impacts what subset of the hypothesis space is considered for search. Thus, this exploration-exploitation trade-off is a key factor in shaping the makeup of the final outcome (Foster & Ford, 2003).

LGM-based planners, including DATAVOYAGER, prefer direct, goal-oriented variables, e.g., preferring *parents' wealth* towards *success in college education*, while de-prioritizing more implicit variables related to urban planning (e.g., *location of schools*). However, while exploration with intrinsic motivators could lead to novel outcomes, it can also sometimes result in false positives (Oudeyer & Kaplan, 2008). How contexts, domains, and the hypothesis space influence the tradeoff between exploration and exploitation remains an open question, which, we argue, is worth considerable research focus (Majumder et al., 2022; Burda et al., 2018).

3.4. Hypothesis Evaluation

Hypothesis Verification. The practical possibility of programmatically verifying a set of hypotheses is a unique feature in data-driven discovery. This encompasses both the proper execution of code as well as the capacity to utilize the appropriate statistical methods and techniques aligned with the high-level research objective (Cai et al., 2023).

The verification of hypotheses can involve (1) the use of tools and (2) code generation. Tools represent a pre-defined set of structured functions, which may be invoked via function-calling by LGMs along with relevant arguments (Pelrine et al., 2023). Code generation, on the other hand, is often unconstrained and can optionally be combined with external tests (Schäfer et al., 2023) and methods such as self-refine (Madaan et al., 2023) in order to minimize hallucination and execution failure.

For example, to verify the hypotheses proposed by the planner, we show DATAVOYAGER's use of independent t-tests to uncover the impact of wealth distributions in two groups on their incarceration probability (Zaw et al., 2016).

```
from scipy import stats
# Perform independent t-tests for the
wealth variables across the two groups
test_results_1985 =
stats.ttest_ind(df[df['ever_jailed'] == 0]
['composite_wealth_1985'],\n
df[df['ever_jailed'] == 1]
['composite_wealth_1985'],\n
equal_var=False)(...)
```

> The results of the independent t-tests for the wealth variables across the two groups (those with and without a criminal record) for the years 1985, 1990, and 1996: (...) T-statistic: 9.7794 (...) *Full Example:* Figure 17

An ideal system must conduct statistical tests (e.g., correlation, regression, multivariate analyses, t-tests or ANOVA for hypothesis testing, etc.), consume execution results, perform analysis to either conclude or re-plan (Prasad et al., 2023) and support usage of domain-specific evaluation toolkits, such as clinical trials (Rotolo et al., 2018) and climate change (Hoffmann et al., 2021).

The complexity of this task arises from the need to support a plethora of analysis tools (see Figure 5) on diverse datasets through unconstrained code generation. Robust verification, further, must be able to analyze execution output and recover from failed initial generation (Ellis et al., 2020). Verification of program output can be enhanced plots, sub-codes, and numerical analyses, yet despite success in math reasoning (Cobbe et al., 2021), LGMs lack multi-modal symbolic understanding (Lu et al., 2023), calling to action the need for improved data experts in systems like DATAVOYAGER.

Continual Learning. Data-driven discovery is an evolving process. With each stage, from hypothesis generation to evaluation, the system collects new insights and successful (or failed) research flows. The system, thus, requires an adaptive learning approach to integrate and understand the changing context and update its understanding of the dataset (Majumder et al., 2023; Shinn et al., 2023) over time.

For example, execution errors while running generated code or failed research pathways provide opportunities for self-refinement and possibly integrating learning into the next instances for more fail-proof planning and execution. Continual learning for data-driven discovery opens up research questions regarding the process of online learning (Majumder et al., 2023; Wang et al., 2023a) involving LGMs and avenues to collect supervision signals for continual fine-tuning (Lin et al., 2022). We argue that how LGMs adapt to novel tools and code at inference time is still an open question and remains critical to data-driven discovery.

3.5. Measurement of Progress

Measuring intermediate progress. Unit tests benchmark intermediate progress in software engineering (Lukasczyk & Fraser, 2022). While a parallel does not exist in ML research, data-driven discovery presents quantitative opportunities to develop internal robust benchmarks for progress evaluation—a property non-existent in almost every discovery system, including DATAVOYAGER. Akin to FunSearch in (Romera-Paredes et al., 2023), we propose to generate a synthetic benchmark with planted hypotheses that are compositionally verifiable for internal evaluation. The infinitely large space of data-generating functions is potent for exploring such data-generation strategies for robust evaluation.

Accommodating human feedback. Autonomous systems can often get stuck, fall into loops, or fail in other unexpected ways. Human feedback corrects errors, prevents unintended paths, and provides necessary interventions ensuring that desired objectives are met. DATAVOYAGER often deviates when fully unsupervised. In the following example, the system focuses on removing multicollinearity despite having a different objective of demographic analysis and having just removed multicollinearity. A user intervention was, thus, necessary.

> **User:** Do not investigate multicollinearity issues. Instead, identify any unique insights or challenges faced by different demographic groups. (...) *Full example:* Figure 16

Despite high degree of natural language fluency, LGM-based systems are often not very proactive. It is desirable for these systems to possess a mixed-initiative ability, thus, optimizing the frequency of asking for human feedback and input (Majumder et al., 2021). Exploring user involvement in the decision-making process raise two questions: (1) *Can we achieve an ideal outcome by enabling users to provide input for tasks like setting low-level objectives or summarizing insights?* (2) *How can we implement effective user intervention during errors or loops to guide the exploration when the system deviates*, as raised in (Lahiri et al., 2022)?

3.6. Knowledge Integration

Interdisciplinary Knowledge Integration. Integrating interdisciplinary knowledge in data-driven discovery enables the interconnection of diverse domains with the high-level research objective, uncovering nuanced associations and insights often overlooked in a single-domain analysis. The challenge lies in internalizing the complexities of different disciplines and recognizing implicit connections, similar to link prediction (Trouillon et al., 2016).

For example, while exploring time preference on BMI (Smith et al., 2005), it could be insightful to assess the role of economic pressure on health outcomes, using cultural anthropology to gauge spending habits, considering psychological factors to understand spending patterns, and proposing strategies for public health intervention and effective urban planning—partially achieved by DATAVOYAGER.

Knowledge Frontiers Support. Knowledge frontiers represent cutting edge scientific exploration and drive groundbreaking discoveries in fields like Machine Learning, gene editing, robotics, and renewable energy (Hassabis, 2002). Enhancing data-driven discovery systems by extending exploration, integrating new methods, and collecting more data can facilitate the investigation of novel scientific domains.

To simulate a knowledge frontier, we accessed a popular language agent repository, Reflexion (Shinn et al., 2023), and modified the experiment design following Majumder et al. (2023). The new experimental data was fed to DATAVOYAGER, which resulted in the following concrete analysis:

> Tasks that are more conceptual or require an understanding of complex systems (e.g., genetics, life stages) seem to be areas where the agent can learn and improve. In contrast, tasks that may involve more practical or hands-on activities (e.g., chemistry mixing, freezing) appear to be more challenging for the agent. (...) *Full example:* Figure 13

We seek to obtain emergent behaviors from curiosity-driven exploration and back-linking to knowledge frontiers (Groth et al., 2021). We raise an open question to automatically search or generate novel datasets (Brickley et al., 2019) and conduct novel exploration with user moderation, leading to data-driven scientific discovery.

3.7. Research Ethics and Fairness

Reproducible Results. Reproducibility stands as a cornerstone of the scientific process (Cao et al., 2023). However, persistent challenges in fields such as economics, psychology, and biomedicine (Camerer et al., 2018; Collaboration, 2015; Fanelli, 2018) in achieving reproducibility call for innovative solutions (Magnusson et al., 2023).

For example, *The Reproducibility Project: Psychology* replicated 100 psychology studies and found only 36% of replications to yield significant results, prompting increased awareness and initiatives to enhance reproducibility across scientific disciplines (Collaboration, 2015). The ideal discovery system should ensure that the undertaken research pathways are reproducible. DATAVOYAGER shows a proof-of-concept for automated, reproducible experiments. However, it can be extended towards automatic documentation and code release, thus further improving transparency.

p-hacking Proof. Manipulating data or analyses to find false significance undermines the scientific process, leading to unreliable findings and subsequent slowdown of progress. For an automated discovery system, this presents a particularly challenging concern and one that can affect its trustworthiness (Wasserstein & Lazar, 2016). p-hacking might involve tweaking variables or testing multiple hypotheses from a dataset until a significant result is found (Dunn, 1961). The data-driven discovery opens up the unique case of evaluating a significant number of hypotheses at the same time, presenting opportunities for unintentional p-hacking. With a large hypotheses space, there is more chance for accidental findings. An ideal data-driven discovery system must perform tests to counter false discoveries (Korthauer et al., 2018) to keep the false discovery rate as low as possible.

4. Limitations of Data-driven Discovery

Hallucinations. LGM-powered data-discovery struggles with output hallucinations, exacerbated by memorization and superposition issues (Elhage et al., 2022) – most susceptible being hypothesis generation, planning, and output comprehension. This undermines the benefits of automation, necessitating external verification and user moderation.

Cost at scale. In high-throughput fields (e.g., computational biology), it is common to test millions of hypotheses (Korthauer et al., 2019). Extensive reliance on these systems for orchestrating experiments can then incur significant computational costs—highlighting the need for integrated cost-benefit analyses into the discovery systems (Agrawal et al., 2023) using, for instance, predictive hazard functions.

Policy misuse. The autonomous discovery system is always at risk of misuse by bad actors to produce a substantial volume of dubious research to fit a particular agenda (Heaven, 2022). For certain disciplines like social science and economics, this could potentially impact policy-making institutions and result in sub-optimal policies and decision-making (Groh et al., 2022).

Legal Implications. Autonomous hypothesis generation and verification, supported by datasets, raise legal challenges around intellectual property rights and authorship (Callison-Burch, 2023) and liability in decision-making processes involving these systems (Farhadi et al., 2023). Defining responsibilities and establishing institutional, legal frameworks to navigate potential suboptimal policies are essential aspects of addressing this challenge.

Underlying Bias. An inherent challenge with the data-driven discovery system involves the potential percolation of bias originating from dual sources—the underlying dataset (Caliskan et al., 2016) and the LGMs (Feng et al., 2023). This introduces the risk of generating hypotheses that reflect and perpetuate existing biases present in the data source being utilized, potentially leading to skewed or unfair insights.

5. Survey on Related Systems

End-to-end Data-driven Discovery Most previous autonomous data-driven discovery systems, such as Bacon (Langley, 1981; Langley et al., 1984; 1983) severely lacked the requisite computational power, restricting their scope with limited discovery of data-driven knowledge. A recent system, CoScientist Boiko et al. (2023), uses LGMs to automate some parts of the workflow; however, it still requires substantial human intervention (e.g., wet lab experiments) for hypothesis verification, thus not qualifying as a fully autonomous discovery system. DataLume (Gu et al., 2023) fully automates the code generation for data transformation and hypothesis verification; however, do not support hy-

	MLAgentBench	CoScientist	Bacon	DataLume	ThoughtSpot	Google AutoML	WolframAlpha*
Objective	Build ML optimal models autonomously	Autonomously plan, execute chemistry experiments	A Production system that discovers empirical laws	Explore data analyst support AI systems can provide	Data plotting, exploration with natural language	Builds optimal black-box model to serve at scale	Automatically analyze data
Comprehensive Data Understanding	Limited to model building	Targeted to Chemical Synthesis	N/A	Data understanding, Transformation	Data Scale only	Data Scale only	Limited to fixed datasets
Hypothesis Generation	N/A	Connect Data and Chemistry Papers, N/A on initial hypothesis	Heuristic-search on data leading to laws or equations	Partially with initial hypothesis	Partially with visualization	N/A	Partially with data analysis
Planning and Orchestrating Research Pathways	Yes, for model performance improvement	Plans for chemical synthesis	N/A, mostly heuristic driven	High-level planning w/o actionable steps	No	No	No
Hypothesis Evaluation	Verification with model performance and LLM efficiency	Conducts physical experiments	Basic heuristic calculations	Verification by interpreting statistical models	Partially with data exploration, visualization	Partially with feature importance	Partially with data analysis
Measurement of Progress	Intrinsic evaluation, but not with human feedback	Accommodates human feedback	N/A	N/A	N/A	Intrinsic model evaluation after training	No
Knowledge Integration	No	Knowledge from web and documents	N/A	Knowledge from LLMs	N/A	N/A	No

Figure 3. **Survey** across several dimensions of a proposed data discovery system for several existing automated and semi-automated data analysis and discovery systems such as: MLAgentBench (Huang et al., 2023), CoScientist (Boiko et al., 2023), Bacon (Langley, 1977), DataLume (Gu et al., 2023), ThoughtSpot (`thoughtspot.com`), Google AutoML (`cloud.google.com/automl`), and Automatic Analysis* from WolframAlpha (`wolframalpha.com/examples/pro-features/data-input`).

pothesis search for complex science workflows. Gil et al. (2022; 2017; 2013) as well as Automatic Analysis in WolframAlpha[7] prototyped various workflows for conducting science in data-driven ways, however, such prototypes never explored the power of LGMs and are only exhibit limited generalizability to datasets and scientific methods.

AutoML AutoML is a workflow of automatically building optimal machine learning and predictive models. AutoML tools exist in scientific packages like Scikit (Feurer et al., 2015) and also in cloud platforms such as Google Cloud Platform[8], Microsoft Azure[9], and Amazon Web Services[10]. Existing AutoML Cloud platform systems primarily perform search over hyperparameters for optimal model development, but they cannot comprehend the semantics of the data and hence cannot help with data-driven hypothesis generation, orchestrating research pathways, and knowledge integration. MLAgentBench (Huang et al., 2023), an evolution of AutoML, performs end-to-end machine learning to benchmark AI research agents. MLAgentBench can plan, evaluate hypotheses, and measure progress, but with a focus on optimizing machine learning models, not on discovering new and novel scientific knowledge.

Automated Data Analysis Automated Data Analysis tools are primarily focused on exploring data under a user-provided query (e.g., "plot sales trends for last 12 months", etc.) and often do not have the capability of searching through the hypotheses space as defined by the data. Spreadsheet tools such as Microsoft Excel and Google Sheets are often part of the scientific workflow as the data analysis tool but show limited automation ability even after coding (Python) support (Monroy, 2023; Google, 2023). Focus on integrating LGMs into data analysis with known workflows (Perlitz et al., 2022; Chakraborty et al., 2024) and code-first data analysis (Santos et al., 2023) increased recently—however, these are limited to small-scale tables and lack abilities such as orchestrating research plans, interpreting results, and knowledge integration.

[7]`https://www.wolframalpha.com/examples/pro-features/data-input`

[8]`cloud.google.com/automl`

[9]`azure.microsoft.com/en-us/products/machine-learning/automatedml/`

[10]`aws.amazon.com/machine-learning/automl/`

6. Conclusion

We argue that ongoing ML research on reasoning, planning, code generation, and tool utilization with LGMs can have a significant influence on advancing and accelerating data-driven discovery. Such systems can transform domains overwhelmed with vast amounts of data, including but not limited to observational social sciences, medicine, astronomy, biology, climate science, computational science, consumer science, and social media analytics.

We posit that the time is ripe for advancing data-driven discovery and that integrating LGMs with tools and user feedback can catalyze notable progress in scientific inquiry. We hope our timely position can increase interest and efforts in developing, debating, and enhancing the vision for an accurate, reliable, and robust system for data-driven discovery. It can help initiate a Cambrian explosion of discovery while promoting speed, reproducibility, and collaboration in scientific research.

Impact Statement

This position paper presents arguments for a goal to advance the field of science by building end-to-end data-driven discovery systems using ML. There are many potential societal consequences of our proposed direction since it involves using large generative models, some of which we cover in our Limitations section, including policy misuse, legal ramifications, and false discovery. On the positive side, our proposed system can advance the rate of discovery, leading to an improved standard of living and social well-being.

Acknowledgments

We sincerely thank Abhijeetsingh Meena, Aryan Prakhar, and Tirth Vora for their engineering and exploration efforts in making DATAVOYAGER. We also thank Peter Jansen, David Wadden, Yoav Golberg, and Daniel Weld for their useful comments. We thank Siddharth Sharma and Siddharth Narayanan for their help with proofreading.

References

Achiam, O. J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J. R., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A. A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D. P., Mu, T., Murati, M., Murk, O., M'ely, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Long, O., O'Keefe, C., Pachocki, J. W., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Pokorny, M., Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M. D., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B. D., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N. A., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report. 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Agarwal, D., Das, R., Khosla, S., and Gangadharaiah, R. Bring your own kg: Self-supervised program synthesis for zero-shot kgqa. *ArXiv*, abs/2311.07850, 2023. URL https://api.semanticscholar.org/CorpusID:265158071.

Agrawal, A., McHale, J., and Oettl, A. Artificial intelligence and scientific discovery: A model of prioritized search. *SSRN Electronic Journal*, 2023. URL https://api.semanticscholar.org/CorpusID:260906716.

Alexander, K. L., Riordan, C., Fennessey, J., and Pallas, A. M. Social background, academic resources, and college graduation: Recent evidence from the national longitudinal survey. *American Journal of Education*, 90(4): 315–333, 1982.

Anderson, C. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7): 16–07, 2008.

Bailis, P., Gan, E., Madden, S., Narayanan, D., Rong, K., and Suri, S. Macrobase: Prioritizing attention in fast data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 541–556, 2017.

Bekhuis, T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical digital libraries*, 3:1–7, 2006.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Bianchini, S., Müller, M., and Pelletier, P. Artificial intelligence in science: An emerging general method of invention. *Research Policy*, 51(10):104604, 2022.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570 – 578, 2023. URL https://api.semanticscholar.org/CorpusID:266432059.

Bowers, S. and Ludäscher, B. An ontology-driven framework for data transformation in scientific workflows. In *International Workshop on Data Integration in the Life Sciences*, pp. 1–16. Springer, 2004.

Brickley, D., Burgess, M., and Noy, N. Google dataset search: Building a search engine for datasets in an open web ecosystem. *The World Wide Web Conference*, 2019. URL https://api.semanticscholar.org/CorpusID:86688027.

Burda, Y., Edwards, H., Storkey, A. J., and Klimov, O. Exploration by random network distillation. *ArXiv*, abs/1810.12894, 2018. URL https://api.semanticscholar.org/CorpusID:53115163.

Cai, T., Wang, X., Ma, T., Chen, X., and Zhou, D. Large language models as tool makers. *ArXiv*, abs/2305.17126, 2023. URL https://api.semanticscholar.org/CorpusID:258947222.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183 – 186, 2016. URL https://api.semanticscholar.org/CorpusID:23163324.

Callison-Burch, C. Understanding generative artificial intelligence and its relationship to copyright. Testimony before The U.S. House of Representatives Judiciary Committee, Subcommittee on Courts, Intellectual Property, and the Internet, May 2023. Hearing on Artificial Intelligence and Intellectual Property: Part I – Interoperability of AI and Copyright Law.

Camerer, C., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., and Wu, H. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2:637 – 644, 2018. URL https://api.semanticscholar.org/CorpusID:52098703.

Cao, H., Dodge, J., Lo, K., McFarland, D. A., and Wang, L. L. The rise of open science: Tracking the evolution and perceived value of data and methods link-sharing practices. *ArXiv*, abs/2310.03193, 2023. URL https://api.semanticscholar.org/CorpusID:263671521.

Chakraborty, A., Banerjee, A., Dasgupta, S., Raturi, V., Soni, A., Gupta, A., Harsola, S., and Subrahmaniam, V. T. Navigator: A gen-ai system for discovery of factual and predictive insights on domain-specific tabular datasets. *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, 2024. URL https://api.semanticscholar.org/CorpusID:266743618.

Chang, J. C., Zhang, A. X., Bragg, J., Head, A., Lo, K., Downey, D., and Weld, D. S. Citesee: Augmenting citations in scientific papers with persistent and personalized historical context. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. URL https://api.semanticscholar.org/CorpusID:256868353.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

Collaboration, O. S. Reproducibility project: Psychology, 2015. URL https://doi.org/10.17605/OSF.IO/EZCUJ.

Dalvi, B., Jansen, P. A., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., and Clark, P. Explaining answers with entailment trees. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID:233297051.

Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64,

1961. URL https://api.semanticscholar.org/CorpusID:122009246.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Elliott, K. C., Cheruvelil, K. S., Montgomery, G. M., and Soranno, P. A. Conceptions of good science in our data-rich world. *BioScience*, 66(10):880–889, 2016.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., and Tenenbaum, J. B. Dreamcoder: growing generalizable, interpretable knowledge with wake–sleep bayesian program learning. *Philosophical Transactions of the Royal Society A*, 381, 2020. URL https://api.semanticscholar.org/CorpusID:219687434.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Fanelli, D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115:2628 – 2631, 2018. URL https://api.semanticscholar.org/CorpusID:4639856.

Farhadi, A., Atkinson, D., Callison-Burch, C., DeCario, N., Dumas, J., Lo, K., and Soldiani, L. AI2's Response to the US Copyright Requence for Comments on Artificial Intelligence and Copyright. US Copyright Office Docket No. 2023-6, 2023. Comment.

Feng, S., Park, C. Y., Liu, Y., and Tsvetkov, Y. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *ArXiv*, abs/2305.08283, 2023. URL https://api.semanticscholar.org/CorpusID:258686693.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. Efficient and robust automated machine learning. *Advances in neural information processing systems*, 28, 2015.

Foster, A. and Ford, N. Serendipity and information seeking: an empirical study. *Journal of documentation*, 59(3):321–340, 2003.

Gil, Y., McWeeney, S. K., and Mason, C. E. Using semantic workflows to disseminate best practices and accelerate discoveries in multi-omic data analysis. In *AAAI Conference on Artificial Intelligence*, 2013. URL https://api.semanticscholar.org/CorpusID:15583030.

Gil, Y., Garijo, D., Ratnakar, V., Mayani, R., Adusumilli, R., Boyce, H., Srivastava, A., and Mallick, P. Towards continuous scientific data analysis and hypothesis evolution. In *AAAI Conference on Artificial Intelligence*, 2017. URL https://api.semanticscholar.org/CorpusID:11269287.

Gil, Y., Khider, D., Osorio, M., Ratnakar, V., Vargas, H., and Garijo, D. Towards capturing scientific reasoning to automate data analysis. 2022. URL https://api.semanticscholar.org/CorpusID:248914202.

Google. Introducing duet ai for google workspace. https://workspace.google.com/blog/product-announcements/duet-ai, 2023. Accessed: 2024-02-18.

Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., and Picard, R. W. Human detection of political speech deepfakes across transcripts, audio, and video. 2022. URL https://api.semanticscholar.org/CorpusID:259342907.

Groth, O., Wulfmeier, M., Vezzani, G., Dasagi, V., Hertweck, T., Hafner, R., Heess, N. M. O., and Riedmiller, M. A. Is curiosity all you need? on the utility of emergent behaviours from curious exploration. *ArXiv*, abs/2109.08603, 2021. URL https://api.semanticscholar.org/CorpusID:237563118.

Gu, K., Grunde-McLaughlin, M., McNutt, A. M., Heer, J., and Althoff, T. How do data analysts respond to ai assistance? a wizard-of-oz study. *ArXiv*, abs/2309.10108, 2023. URL https://api.semanticscholar.org/CorpusID:262054482.

Hassabis, D. Using ai to accelerate scientific discovery, 2002. URL https://www.youtube.com/watch?v=jocWJiztxYA&ab_channel=InstituteforEthicsinAIOxford.

Heaven, W. D. Why meta's latest large language model survived only three days online. *MIT Technology Review. Last accessed December*, 15:2022, 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. X., and Steinhardt, J. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Hennig, P. and Schuler, C. J. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

Hoffmann, C. G., Kiladis, G. N., Gehne, M., and von Savigny, C. A python package to calculate the olr-based index of the madden-julian-oscillation (omi) in climate science and weather forecasting. *Journal of Open Research Software*, 2021. URL https://api.semanticscholar.org/CorpusID:236586655.

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J. C., and Wang, H. Large language models for software engineering: A systematic literature review. *ArXiv*, abs/2308.10620, 2023. URL https://api.semanticscholar.org/CorpusID:261048648.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. Benchmarking large language models as ai research agents. *ArXiv*, abs/2310.03302, 2023. URL https://api.semanticscholar.org/CorpusID:263671541.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011.

Kambhampati, S., Valmeekam, K., Guan, L., Stechly, K., Verma, M., Bhambri, S., Saldyt, L., and Murthy, A. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

Kang, D., Emmons, J., Abuzaid, F., Bailis, P. D., and Zaharia, M. A. Noscope: Optimizing deep cnn-based queries over video streams at scale. *Proc. VLDB Endow.*, 10:1586–1597, 2017. URL https://api.semanticscholar.org/CorpusID:20732104.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., and Sabharwal, A. Decomposed prompting: A modular approach for solving complex tasks. *ArXiv*, abs/2210.02406, 2022. URL https://api.semanticscholar.org/CorpusID:252715485.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20, 2019. doi: 10.1186/s13059-019-1716-1. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1716-1.

Korthauer, K. D., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C. J., Alm, E. J., and Hicks, S. C. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20, 2018. URL https://api.semanticscholar.org/CorpusID:91264977.

Lahiri, S. K., Naik, A., Sakkas, G., Choudhury, P., von Veh, C., Musuvathi, M., Inala, J. P., Wang, C., and Gao, J. Interactive code generation via test-driven user-intent formalization. *ArXiv*, abs/2208.05950, 2022. URL https://api.semanticscholar.org/CorpusID:251492970.

Langley, P. Bacon: A production system that discovers empirical laws. In *International Joint Conference on Artificial Intelligence*, 1977. URL https://api.semanticscholar.org/CorpusID:2320342.

Langley, P. Data-driven discovery of physical laws. *Cogn. Sci.*, 5:31–54, 1981. URL https://api.semanticscholar.org/CorpusID:39694251.

Langley, P., Bradshaw, G. L., and Simon, H. A. Rediscovering chemistry with the bacon system. 1983. URL https://api.semanticscholar.org/CorpusID:118714327.

Langley, P., Zytkow, J. M., Simon, H. A., and Bradshaw, G. L. The search for regularity: Four aspects of scientific discovery. 1984. URL https://api.semanticscholar.org/CorpusID:3155192.

Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., Groves-Kirkby, N., Mihajlovic, A., DiGiovanna, J., Srdic, M., et al. The cancer genomics cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer research*, 77(21):e3–e6, 2017.

LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

Leeman, J., Liu, Y., Stiles, J., Lee, S., Bhatt, P., Schoop, L., and Palgrave, R. Challenges in high-throughput inorganic material prediction and autonomous synthesis. 2024.

Li, Y., Choi, D. H., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Tom, Eccles, Keeling, J., Gimeno, F., Lago, A. D., Hubert, T., Choy, P., de, C., d'Autume, M., Babuschkin, I., Chen, X., Huang, P.-S., Welbl, J., Gowal, S., Alexey, Cherepanov, Molloy, J., Mankowitz, D. J., Robson, E. S., Kohli, P., de, N., Freitas, Kavukcuoglu, K., and Vinyals, O. Competition-level code generation with alphacode. *Science*, 378:1092 – 1097, 2022. URL https://api.semanticscholar.org/CorpusID:246527904.

Lin, B. Y., Wang, S. I., Lin, X. V., Jia, R., Xiao, L., Ren, X., and tau Yih, W. On continual model refinement in out-of-distribution data streams. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:248512744.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023a. URL https://api.semanticscholar.org/CorpusID:258179774.

Liu, J., Xia, C., Wang, Y., and Zhang, L. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *ArXiv*, abs/2305.01210, 2023b. URL https://api.semanticscholar.org/CorpusID:258437095.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., and Tang, J. Agentbench: Evaluating llms as agents. *ArXiv*, abs/2308.03688, 2023c. URL https://api.semanticscholar.org/CorpusID:260682249.

Lu, P., Bansal, H., Xia, T., Liu, J., yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. 2023. URL https://api.semanticscholar.org/CorpusID:264491155.

Lukasczyk, S. and Fraser, G. Pynguin: Automated unit test generation for python. *2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pp. 168–172, 2022. URL https://api.semanticscholar.org/CorpusID:246706202.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Welleck, S., Majumder, B. P., Gupta, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651, 2023. URL https://api.semanticscholar.org/CorpusID:257900871.

Magnusson, I. H., Smith, N. A., and Dodge, J. Reproducibility in nlp: What have we learned from the checklist? In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:259187997.

Majumder, B. P., Rao, S., Galley, M., and McAuley, J. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *North American Chapter of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:233231257.

Majumder, B. P., Jhamtani, H., Berg-Kirkpatrick, T., and McAuley, J. Achieving conversational goals with unsupervised post-hoc knowledge injection. *ArXiv*, abs/2203.11399, 2022. URL https://api.semanticscholar.org/CorpusID:247547046.

Majumder, B. P., Dalvi, B., Jansen, P., Tafjord, O., Tandon, N., Zhang, L., Callison-Burch, C., and Clark, P. Clin: A continually learning language agent for rapid task adaptation and generalization. *ArXiv*, abs/2310.10134, 2023. URL https://api.semanticscholar.org/CorpusID:264146262.

Monroy, D. Introducing copilot support for python in excel: Advanced data analysis using natural language. https://techcommunity.microsoft.com/t5/excel-blog/introducing-copilot-support-for-python-in-excel-advanced-data/ba-p/3928120, 2023. Accessed: 2024-02-18.

Newell, A. and Simon, H. A. Computer science as empirical inquiry: symbols and search. *Commun. ACM*, 19(3):113–126, mar 1976. ISSN 0001-0782. doi: 10.1145/360018.360022. URL https://doi.org/10.1145/360018.360022.

Oudeyer, P.-Y. and Kaplan, F. How can we define intrinsic motivation? 2008. URL https://api.semanticscholar.org/CorpusID:14217330.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2): 265–286, 2007.

Palani, S., Naik, A., Downey, D., Zhang, A. X., Bragg, J., and Chang, J. C. Relatedly: Scaffolding literature reviews with existing related work sections. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. URL https://api.semanticscholar.org/CorpusID:256846632.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Pelrine, K., Taufeeque, M., Zajkac, M., McLean, E., and Gleave, A. Exploiting novel gpt-4 apis. *ArXiv*, abs/2312.14302, 2023. URL https://api.semanticscholar.org/CorpusID:266521205.

Perlitz, Y., Sheinwald, D., Slonim, N., and Shmueli-Scheuer, M. nbiig: A neural bi insights generation system for table reporting. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:253397856.

Prasad, A., Koller, A., Hartmann, M., Clark, P., Sabharwal, A., Bansal, M., and Khot, T. Adapt: As-needed decomposition and planning with language models. *ArXiv*, abs/2311.05772, 2023. URL https://api.semanticscholar.org/CorpusID:265128575.

Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., Wang, B., Kim, Y., Choi, Y., Dziri, N., and Ren, X. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *ArXiv*, abs/2310.08559, 2023. URL https://api.semanticscholar.org/CorpusID:263909078.

Ramakrishnan, N. and Grama, A. Y. Data mining: From serendipity to science. *Computer*, 32(8):34–37, 1999.

Ristoski, P. and Paulheim, H. Semantic web in data mining and knowledge discovery: A comprehensive survey. *J. Web Semant.*, 36:1–22, 2016. URL https://api.semanticscholar.org/CorpusID:42846121.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., Fawzi, A., Grochow, J., Lodi, A., Mouret, J.-B., Ringer, T., and Yu, T. Mathematical discoveries from program search with large language models. *Nature*, 625:468 – 475, 2023. URL https://api.semanticscholar.org/CorpusID:266223700.

Rotolo, F., Paoletti, X., and Michiels, S. surrosurv: An r package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. *Computer methods and programs in biomedicine*, 155:189–198, 2018. URL https://api.semanticscholar.org/CorpusID:3480478.

Santos, M., Clemente, F., and Abshire, C. Pandas-profiling now supports apache spark. https://www.databricks.com/blog/2023/04/03/pandas-profiling-now-supports-apache-spark.html, 2023. Accessed: 2024-02-18.

Schäfer, M., Nadi, S., Eghbali, A., and Tip, F. An empirical evaluation of using large language models for automated unit test generation. *IEEE Transactions on Software Engineering*, 50:85–105, 2023. URL https://api.semanticscholar.org/CorpusID:256827098.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Yacmpz84TH.

Sharma, A., Li, X., Guan, H., Sun, G., Zhang, L., Wang, L., Wu, K., Cao, L., Zhu, E., Sim, A., Wu, T., and Zou, J. Automatic data transformation using large language model - an experimental study on building energy data. *2023 IEEE International Conference on Big Data (BigData)*, pp. 1824–1834, 2023. URL https://api.semanticscholar.org/CorpusID:261530167.

Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. 2023. URL https://api.semanticscholar.org/CorpusID:258833055.

Smith, P. K., Bogin, B., and Bishai, D. Are time preference and body mass index associated?: Evidence from the national longitudinal survey of youth. *Economics & Human Biology*, 3(2):259–270, 2005.

Stanley, K. O., Lehman, J., and Soros, L. Open-endedness: The last grand challenge you've never heard of. *While open-endedness could be a force for discovering intelligence, it could also be a component of AI itself*, 2017.

Sun, R., Arik, S. Ö., Nakhost, H., Dai, H., Sinha, R., Yin, P., and Pfister, T. Sql-palm: Improved large language model adaptation for text-to-sql. *ArXiv*, abs/2306.00739, 2023. URL https://api.semanticscholar.org/CorpusID:258999853.

Swanson, D. R. Undiscovered public knowledge. *The Library Quarterly*, 56:103–118, 1986. URL https:

//api.semanticscholar.org/CorpusID:144270735.

Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*. Random House Group, 2007. ISBN 1400063515.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. *ArXiv*, abs/1606.06357, 2016. URL https://api.semanticscholar.org/CorpusID:15150247.

Vaithilingam, P., Zhang, T., and Glassman, E. L. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022. URL https://api.semanticscholar.org/CorpusID:247255943.

Valmeekam, K., Olmo, A., Sreedharan, S., and Kambhampati, S. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. 2022. URL https://api.semanticscholar.org/CorpusID:249889477.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L. J., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *ArXiv*, abs/2305.16291, 2023a. URL https://api.semanticscholar.org/CorpusID:258887849.

Wang, Q., Downey, D., Ji, H., and Hope, T. Learning to generate novel scientific directions with contextualized literature-based discovery. *ArXiv*, abs/2305.14259, 2023b. URL https://api.semanticscholar.org/CorpusID:258841365.

Wasserstein, R. and Lazar, N. A. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70:129 – 133, 2016. URL https://api.semanticscholar.org/CorpusID:124084622.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Weiss, S. T. and Ware, J. H. Overview of issues in the longitudinal analysis of respiratory data. *American journal of respiratory and critical care medicine*, 154 6 Pt 2:S208–11, 1996. URL https://api.semanticscholar.org/CorpusID:45049299.

Weitzman, M. Optimal search for the best alternative. *Econometrica*, 47:641–654, 1978. URL https://api.semanticscholar.org/CorpusID:32530881.

Whitley, L. D. Fundamental principles of deception in genetic search. In *Foundations of genetic algorithms*, volume 1, pp. 221–241. Elsevier, 1991.

Wolf, Y., Wies, N., Levine, Y., and Shashua, A. Fundamental limitations of alignment in large language models. *ArXiv*, abs/2304.11082, 2023. URL https://api.semanticscholar.org/CorpusID:258291526.

Zaw, K., Hamilton, D., and Darity, W. A. J. Race, wealth and incarceration: Results from the national longitudinal survey of youth. *Race and Social Problems*, 8:103–115, 2016. URL https://api.semanticscholar.org/CorpusID:13709779.

Zhang, J., Lehman, J., Stanley, K., and Clune, J. Omni: Open-endedness via models of human notions of interestingness. *arXiv preprint arXiv:2306.01711*, 2023.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., and Sap, M. Sotopia: Interactive evaluation for social intelligence in language agents. *ArXiv*, abs/2310.11667, 2023. URL https://api.semanticscholar.org/CorpusID:264289186.

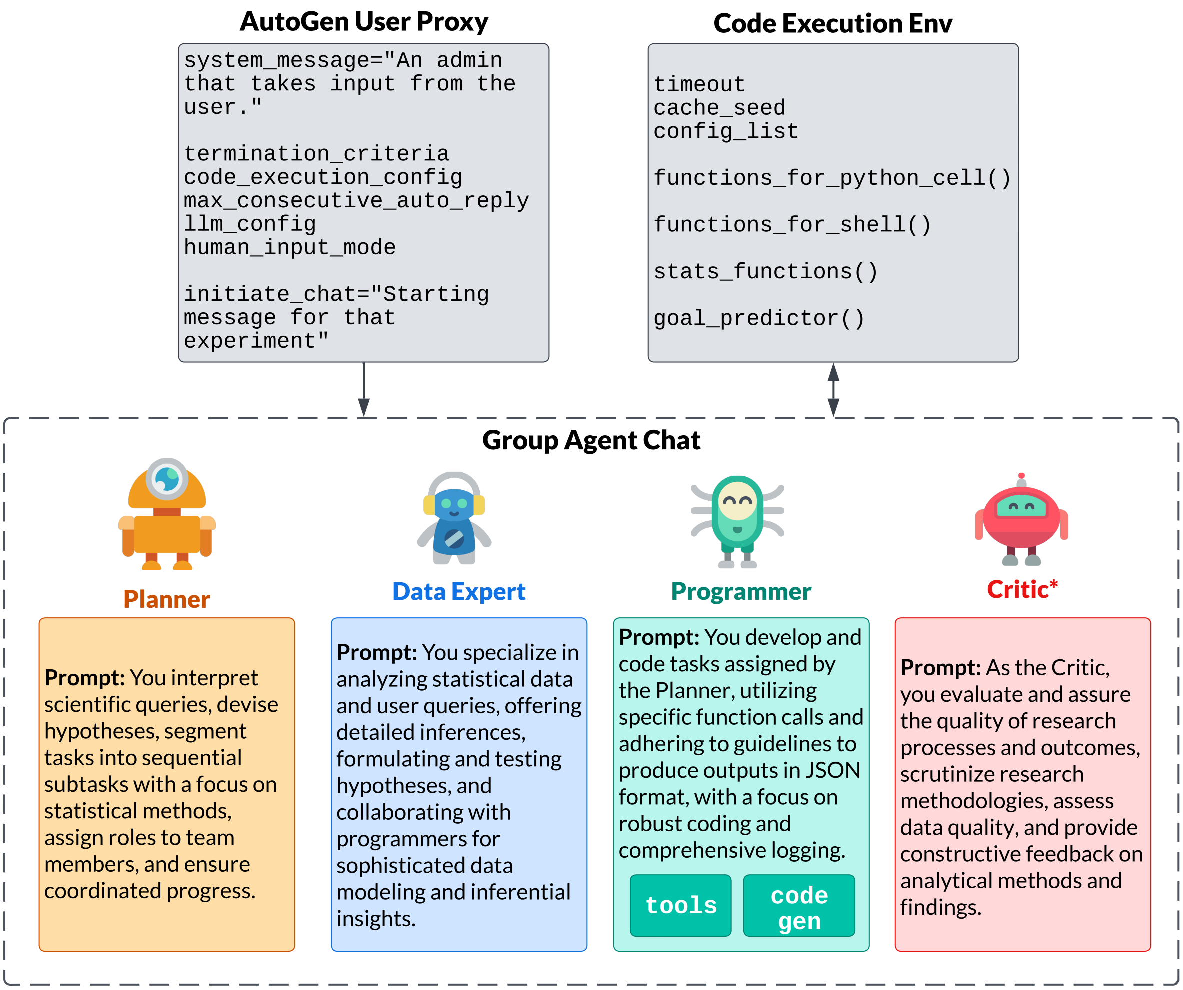

Figure 4. Agent Structure for DATAVOYAGER. Group Agent Chat has AutoGen agents that communicate with each other. The User Proxy links the user with the agents to share data, feedback, and goals. Code Execution Environment has access to structured functions and code generation methods that can be called depending on the context.

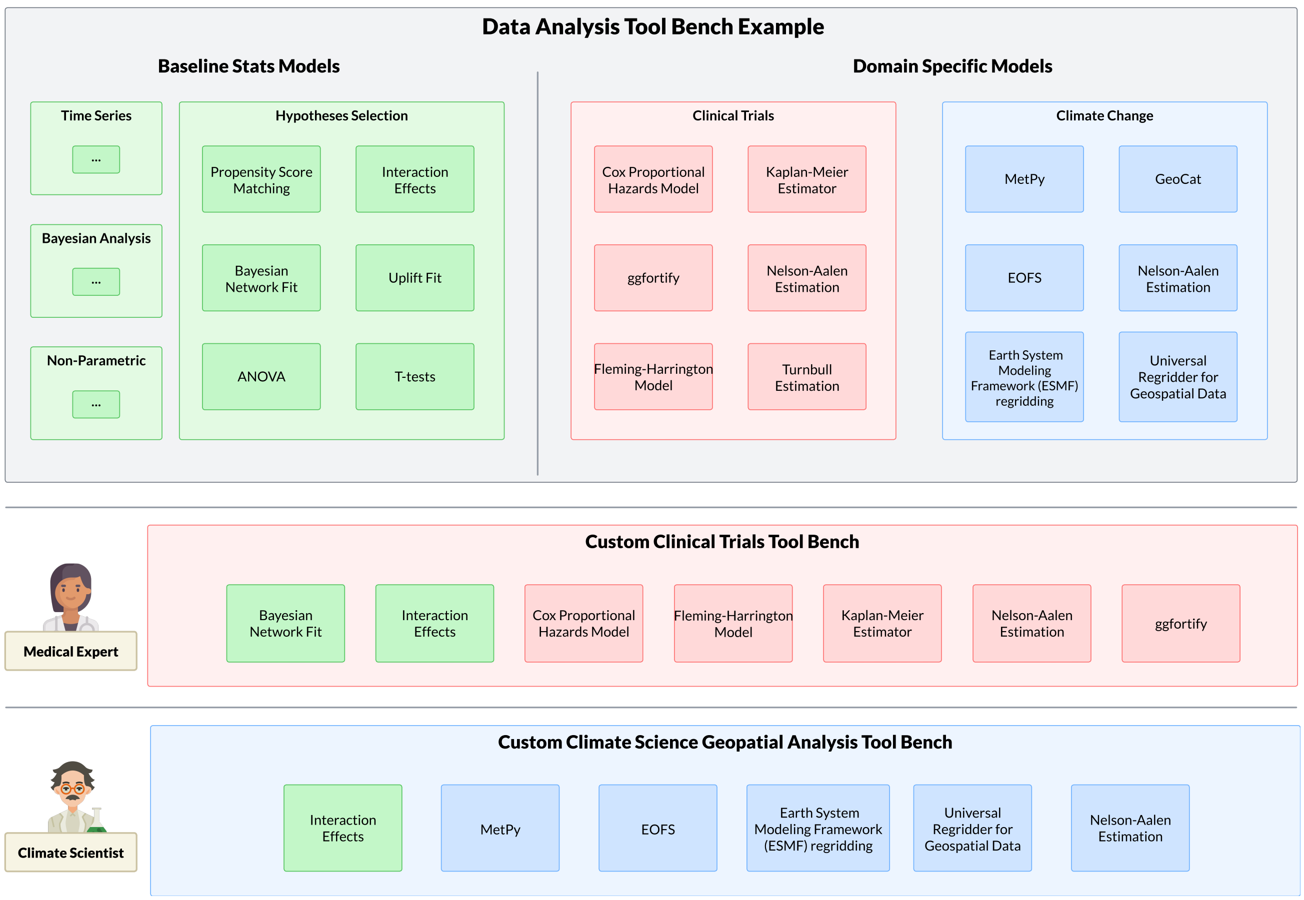

Figure 5. Data Analysis Tool Bench that can be structured inside DATAVOYAGER to enable discovery in a wide range of scientific domains.

User Proxy (to chat_manager):

Examine the effects of Time Preference on Individuals

Data Modelling Analyst Expert (to chat_manager):

To examine the effects of time preference on individuals, we need to understand the variables in the dataset that relate to time preference. In the provided dataset, the variables `DISSAVED` and `SAMESAVE` seem to be related to time preference. Typically, time preference refers to the extent to which individuals prefer to receive goods or services sooner rather than later.

Here's what we can infer about the variables:

- `DISSAVED`: This variable might indicate individuals who have a present bias, preferring current consumption over future savings. A value of 1 could indicate individuals who have dissaved, while a value of 0 indicates individuals who have not dissaved.
- `SAMESAVE`: This variable might indicate individuals who maintain their savings, neither dissaving nor actively saving more. A value of 1 could indicate individuals who have the same savings, while a value of 0 indicates individuals who do not.

To examine the effects of these time preference indicators on individuals, we can look at their relationships with other variables in the dataset, such as `AGE`, `INCOME`, `BMI`, and demographic variables like `BLACK`, `HISPANIC`, `WHITE`, `MALE`, and `FEMALE`.

We can perform statistical analyses such as correlation analysis, regression analysis, or comparison of means to understand these relationships. For example, we might hypothesize that individuals with a higher tendency to dissave (`DISSAVED`=1) have different income levels or BMI values compared to those who do not dissave.

Let's start by performing a descriptive statistical analysis to summarize the data, followed by a correlation analysis to see the relationships between time preference variables and other variables in the dataset.

Team member, please execute the following Python code to perform the descriptive statistical analysis and correlation analysis:

```
# Descriptive statistics for the dataset
descriptive_stats = df.describe()

# Correlation analysis
correlation_matrix = df.corr()

(descriptive_stats, correlation_matrix)
```

Figure 6. **Background:** Data from National Longitudinal Survey of Youth along with question on relation between time preference & BMI was fed into DATAVOYAGER; it is a question studied in (Smith et al., 2005). **This figure:** Data Understanding - In response to a high-level objective, the system demonstrates the need of understanding the variables before initiating statistical analysis. Moreover, it selects relevant Time Preference variables from the data and infers them (highlighted in green).

User Proxy (to chat_manager):

Examine the potential link between Time Preference and Health variable based on demographics

Data Modelling Analyst Expert (to chat_manager):

To examine the potential link between time preference and health, as indicated by the Body Mass Index (BMI), and how this relationship may vary based on demographics, we can perform subgroup analyses and interaction term analyses within our regression models. This will allow us to see if the relationship between time preference (as indicated by `DISSAVED` and `SAMESAVE`) and BMI differs across different demographic groups such as race and gender.

Here are the steps we can take to explore this link:

1. **Subgroup Analysis**: We can stratify the dataset by demographic variables (e.g., `BLACK`, `HISPANIC`, `WHITE`, `MALE`, `FEMALE`) and perform separate regression analyses for each subgroup to see if the relationship between time preference and BMI differs among these groups.
2. **Interaction Term Analysis**: We can add interaction terms between BMI and demographic variables to our regression models to directly test whether the effect of BMI on time preference behaviors varies by demographic group.

Let's start by adding interaction terms to our logistic regression models to examine the potential link between time preference and BMI across different demographic groups.

Team member, please execute the following Python code to perform logistic regression with interaction terms for `DISSAVED` and `SAMESAVE`:

```
# Create interaction terms between BMI and demographic variables
df['BMI_BLACK'] = df['BMI'] * df['BLACK']
df['BMI_HISPANIC'] = df['BMI'] * df['HISPANIC']
df['BMI_WHITE'] = df['BMI'] * df['WHITE']
df['BMI_MALE'] = df['BMI'] * df['MALE']
df['BMI_FEMALE'] = df['BMI'] * df['FEMALE']

# Logistic regression with interaction terms for DISSAVED
X_dissaved_interaction = df[['AGE', 'INCOME', 'BMI', 'BLACK', 'HISPANIC',
'WHITE', 'MALE', 'FEMALE', 'BMI_BLACK', 'BMI_HISPANIC', 'BMI_WHITE',
'BMI_MALE', 'BMI_FEMALE']]
y_dissaved = df['DISSAVED']

logit_model_dissaved_interaction = sm.Logit(y_dissaved,
sm.add_constant(X_dissaved_interaction)).fit()

# Logistic regression with interaction terms for SAMESAVE
X_samesave_interaction = df[['AGE', 'INCOME', 'BMI', 'BLACK', 'HISPANIC',
'WHITE', 'MALE', 'FEMALE', 'BMI_BLACK', 'BMI_HISPANIC', 'BMI_WHITE',
'BMI_MALE', 'BMI_FEMALE']]
y_samesave = df['SAMESAVE']

logit_model_samesave_interaction = sm.Logit(y_samesave,
sm.add_constant(X_samesave_interaction)).fit()

(logit_model_dissaved_interaction.summary(),
logit_model_samesave_interaction.summary())
```

Figure 7. **Background:** Data from National Longitudinal Survey of Youth along with question on relation between time preference & BMI was fed into DATAVOYAGER; it is a question studied in (Smith et al., 2005). **This figure:** Data Transformation - The system generated insights from the results of Logistic Regression with L1 regularization. As a response to user input, the system showcases data transformation ability by creating new interaction terms in logistic regression models (as demonstrated in the code snippet), exploring the link between time preference and BMI across diverse demographic groups.

Data Modelling Analyst Expert (to chat_manager):

The OLS regression results, which show varying degrees of association between financial behaviors and BMI across different demographic groups, can be further explored through an interdisciplinary lens to generate broader insights:

1. **Economics and Behavioral Economics:**
   - The significant relationship between financial behaviors and BMI in certain groups, such as Hispanic Males, may reflect underlying economic pressures or the influence of economic policies on health outcomes.
   - Behavioral economics can help understand how financial decision-making, influenced by cognitive biases and social norms, might relate to health behaviors and BMI.
2. **Sociology and Cultural Anthropology:**
   - The differences in the significance and strength of associations across demographic groups suggest that cultural norms and societal roles related to gender and race may influence both financial behaviors and health outcomes.
   - Cultural anthropology can provide insights into how cultural practices around food, body image, and money management differ across these groups and contribute to the observed patterns.
3. **Psychology and Social Psychology:**
   - The role of 'SAMESAVE' in predicting BMI may be tied to social comparison and the desire for conformity, which are well-studied concepts in social psychology.
   - Psychological factors such as stress, which can be related to financial insecurity ('DISSAVED'), may have different impacts on health behaviors and BMI across demographic groups.
4. **Public Health and Health Promotion:**
   - The findings could inform public health strategies that address both financial well-being and healthy lifestyle choices, particularly in groups where a stronger association is observed.
   - Health promotion interventions could be tailored to address the specific needs and cultural contexts of each demographic group to improve BMI outcomes.
5. **Urban Planning and Policy Studies:**
   - Urban planning can consider how neighborhood design influences both economic opportunities and access to healthy food and recreational spaces, which can impact BMI.
   - Policy studies can explore how policies aimed at reducing economic inequality and improving financial security might also have beneficial effects on public health.

Interdisciplinary Hypotheses:

1. **Integrated Health and Financial Interventions:** Interventions that simultaneously address financial literacy and health education may be more effective in improving health outcomes, particularly in groups where financial behaviors are significantly associated with BMI.
2. **Cultural Competence in Health Promotion:** Health promotion programs that are culturally competent and consider the unique social and economic contexts of each demographic group may lead to better engagement and outcomes.
3. **Social Support Networks:** The role of social support networks in both financial and health behaviors could be a critical factor in understanding the relationship between 'SAMESAVE' and BMI, suggesting that strengthening these networks might improve health outcomes.

By considering these interdisciplinary insights, researchers and policymakers can develop a more nuanced understanding of the factors that influence BMI and design more effective interventions that are sensitive to the complexities of race, gender, and cultural differences.

Figure 8. **Background:** Data from National Longitudinal Survey of Youth along with question on relation between time preference & BMI was fed into DATAVOYAGER; it is a question studied in (Smith et al., 2005). **This figure:** Interdisciplinary Knowledge Integration - The system extracts insights from BMI data, generates insights (highlighted in yellow) from the lens of different disciplines and integrates them into different interdisciplinary hypotheses for further exploration (highlighted in green).

Data Modelling Analyst Expert (to chat_manager):

The Generalized Linear Model (GLM) regression results with a Gaussian family and identity link function provide the following insights:

- **Model Fit**: The GLM results are quite similar to the OLS regression results, which is expected given that the Gaussian family with an identity link function is equivalent to the OLS model. The coefficients and their significance levels are nearly identical to those found in the OLS model with interaction terms.
- **Interaction Effects**:
  - The interaction between "DISSAVED" and "GENDER_MALE" remains significant, with a coefficient of 0.5259 ($p < 0.000$). This reinforces the finding that the relationship between dissaving and BMI is more pronounced for males.
  - The interaction between "DISSAVED" and "RACE_WHITE" is also significant, with a coefficient of -0.5748 ($p < 0.001$), suggesting that White individuals who have dissaved have a lower BMI than what would be predicted by the main effects alone.
  - The interaction between "SAMESAVE" and "RACE_HISPANIC" is significant, with a coefficient of 0.6251 ($p < 0.002$), indicating that Hispanics who maintain the same level of savings have a higher BMI.
- **Main Effects**:
  - "AGE" and "INCOME" continue to show significant associations with BMI, with age positively associated and income negatively associated with BMI.
  - "GENDER_MALE" has a significant positive association with BMI, indicating that males have a higher BMI than females.
  - "RACE_WHITE" has a significant negative association with BMI compared to the reference race group.
- **Model Diagnostics**:
  - The Pearson chi-squared statistic and the Pseudo R-squared value are provided as measures of model fit. The Pseudo R-squared value is slightly higher than the R-squared from the OLS model, but it still indicates that the model explains a relatively small portion of the variability in BMI.
- **New Insights**:
  - The GLM confirms the findings from the OLS model regarding the interactions between time preference and demographic factors. This suggests that the observed relationships are robust across different modeling approaches.
  - The consistency of the results across both OLS and GLM models strengthens the confidence in the identified relationships between time preference, demographic factors, and BMI.
- **Hypotheses for Future Research**:
  - The significant interactions suggest that future research should explore the mechanisms by which gender and race influence the relationship between time preference and health outcomes.
  - The role of socioeconomic status, as indicated by income, in health behaviors and outcomes should be further investigated.
  - Additional factors not included in the current model, such as diet, exercise, and genetic predispositions, should be considered to better understand the determinants of BMI.

In conclusion, the sophisticated analysis using a GLM confirms the insights gained from the OLS model with interaction terms. The relationships between time preference, demographic factors, and BMI are complex and influenced by a variety of factors. These findings can guide future research to explore the underlying causes and implications of these relationships for public health interventions and policies.

Figure 9. **Background:** Data from National Longitudinal Survey of Youth along with question on relation between time preference & BMI was fed into DATAVOYAGER; it is a question studied in (Smith et al., 2005). **This figure:** Hypothesis Verification - When the user prompted to perform sophisticated analysis to uncover new insights, the system generates new insights utilizing the Generalized Linear Model (highlighted in blue) that confirms the results from the previous OLS analysis (highlighted in green).

User Proxy (to chat_manager):

list out possible analyses that we can perform to generate interesting insights from the data

Data Modelling Analyst Expert (to chat_manager):

Several analyses can be performed on this dataset to generate insights into the learning process of the agents and the difficulty of the tasks. Here are some possible analyses:

1. **Learning Progression Analysis**:
   - Track the improvement in scores from episode 0 to the maximum score across different task variants.
   - Analyze the number of episodes it takes for agents to reach their maximum score.
2. **Task Difficulty Assessment**:
   - Compare the maximum scores and the number of episodes required across different task variants to assess relative difficulty.
   - Identify task variants where agents consistently fail to improve beyond the initial score.
3. **Learning Efficiency**:
   - Calculate the average improvement in score per episode to gauge learning efficiency.
   - Compare learning efficiency across different task variants.
4. **Episode Analysis**:
   - Investigate the distribution of `max_episode_number` to understand how long agents tend to engage with the tasks.
   - Analyze the correlation between the number of episodes and the score improvement.
5. **Score Distribution Analysis**:
   - Examine the distribution of `episode_0_score` and `max_score` to identify patterns in initial performance and maximum potential.
   - Look for outliers or anomalies in score distributions that might indicate unusual learning patterns or data entry errors.
6. **Binary Outcome Analysis**:
   - Analyze the proportion of task variants where `max_episode_greater_than_0` is 1 versus 0 to understand how often agents progress past the initial episode.
   - Investigate the relationship between binary outcomes and task difficulty.
7. **Comparative Analysis**:
   - Compare the learning outcomes based on all state changes, final state, and final score to determine which metric is the most indicative of agent performance.
   - Assess whether certain task variants show more significant differences between these metrics.
8. **Regression Analysis**:
   - Perform regression analysis to predict the maximum score based on the initial score and the number of episodes.
   - Identify significant predictors of learning success.
9. **Cluster Analysis**:
   - Use clustering techniques to group task variants based on learning outcomes and identify common characteristics within each cluster.
10. **Time-Series Analysis**:
    - If the data can be ordered chronologically, perform time-series analysis to identify trends over time in the learning process.
11. **Hypothesis Testing**:
    - Formulate and test hypotheses regarding the impact of different task variants on learning outcomes.
    - Use statistical tests to determine if observed differences in scores are statistically significant.

Each of these analyses can help to uncover different aspects of the learning process, the effectiveness of the agents, and the nature of the tasks themselves. The insights gained can be used to improve the design of the learning algorithms, adjust the difficulty of the tasks, or tailor the learning environment to better suit the agents' capabilities.

Figure 10. **Background:** Experimental data of running methods using a popular agent-based repo Reflexion (`https://github.com/noahshinn/reflexion`) and (Majumder et al., 2023) was fed to DataVoyager. **This figure:** Knowledge Frontiers Support - Data Expert suggested interesting list of analyses to find new insights. New analyses (highlighted in green) were created with limited context on the data just based on variable description. Cluster analysis (highlighted in blue) leads to novel insights in the agent literature.

User Proxy (to chat_manager):

Proceed with learning progression analysis and generate detailed insights from its results

Planner (to chat_manager):

To perform a Learning Progression Analysis, we will follow these steps:

1. **Calculate Improvement**: Compute the improvement in scores from episode 0 to the maximum score for each task variant and for each of the three categories (all state changes, final state, and final score).
2. **Episode Analysis**: Analyze the number of episodes it takes for agents to reach their maximum score and how this varies across task variants.
3. **Aggregate Statistics**: Calculate aggregate statistics such as mean, median, standard deviation, and quartiles for score improvements and episode numbers.
4. **Visualizations**: Create visualizations such as line plots, histograms, or box plots to illustrate the distribution of score improvements and episode numbers.
5. **Correlation Analysis**: Determine if there is a correlation between the initial score and the improvement, as well as between the number of episodes and the improvement.
6. **Task Variant Comparison**: Compare the learning progression across different task variants to identify which ones show the most and least improvement.
7. **Insight Generation**: Based on the results of the above analyses, generate detailed insights regarding the learning progression of agents.

Let's start by assigning the first subtask to calculate the improvement in scores and analyze the number of episodes for each category.

Subtask 1: Calculate Improvement and Episode Analysis

- Calculate the difference between the maximum score and the episode 0 score for all state changes, final state, and final score.
- Calculate the average number of episodes taken to reach the maximum score for each category.
- Prepare the data for aggregate statistics and visualizations.

I will assign this subtask to the data analyst team member.

Arguments for function call python:

```
{"cell":"# Subtask 1: Calculate Improvement and Episode Analysis\n\n#
Calculate the score improvement for each category\nimprovement_columns =
[col for col in df.columns if
'difference_between_max_score_and_episode_0_score' in col]\ndf_improvement
= df[improvement_columns].describe()\n\n# Calculate the average number of
episodes for each category\nepisode_columns = [col for col in df.columns if
'max_episode_number' in col]\ndf_episodes =
df[episode_columns].mean()\n\n(df_improvement, df_episodes)"}
```

Figure 11. **Background:** Experimental data of running methods using a popular agent-based repo Reflexion (https://github.com/noahshinn/reflexion) and (Majumder et al., 2023) was fed to DATAVOYAGER. **This figure:** Multi-Step Planning - The system understood the variables and carved out the steps that need to be performed to draw interesting insights (highlighted in blue). The Planner created an excellent plan by breaking the objective into subtasks to carry-out learning progression analysis. It then assigned the subtask to a team member (highlighted in yellow).

User Proxy (to chat_manager):

Based upon these insights, generate hypotheses and perform sophisticated statistical analysis to uncover more complex underlying patterns

Data Modelling Analyst Expert (to chat_manager):

Based on the insights generated, we can formulate several hypotheses to investigate more complex underlying patterns in the data. Here are some hypotheses we can test:

1. **Hypothesis 1: Task Difficulty**
   - H0: There is no significant difference in initial performance across tasks.
   - H1: There is a significant difference in initial performance across tasks.
   - Test: ANOVA to compare the means of `episode_0_score_for_all_state_changes` across different `task_name` groups.
2. **Hypothesis 2: Learning Effect**
   - H0: There is no significant improvement from the initial to the maximum score for all state changes across tasks.
   - H1: There is significant improvement from the initial to the maximum score for all state changes across tasks.
   - Test: Paired t-test or Wilcoxon signed-rank test comparing `episode_0_score_for_all_state_changes` and `max_score_for_all_state_changes` within each task.
3. **Hypothesis 3: Consistency in Performance**
   - H0: The number of episodes greater than 0 is consistent across tasks.
   - H1: The number of episodes greater than 0 varies significantly across tasks.
   - Test: Chi-square test for independence to see if the distribution of `max_episode_greater_than_0_for_all_state_changes` is associated with `task_name`.
4. **Hypothesis 4: Final State Achievement**
   - H0: The final state achievement is consistent across tasks.
   - H1: The final state achievement varies significantly across tasks.
   - Test: ANOVA to compare the means of `max_score_for_final_state` across different `task_name` groups.
5. **Hypothesis 5: Efficiency in Task Completion**
   - H0: The efficiency in reaching the final state is consistent across tasks.
   - H1: The efficiency in reaching the final state varies significantly across tasks.
   - Test: Kruskal-Wallis H-test to compare the distribution of `max_episode_number_for_final_state` across different `task_name` groups, as the data may not be normally distributed.

To perform these statistical tests, we will need to execute Python code for each hypothesis test. Let's start with Hypothesis 1, testing for task difficulty using ANOVA.

Subtask: Perform ANOVA to test Hypothesis 1.

```python
import pandas as pd
from scipy import stats

# Load the data
df = pd.read_csv('./CLIN for OSOD - Transformed.csv')

# Perform ANOVA on 'episode_0_score_for_all_state_changes' across different
'task_name' groups
anova_results =
stats.f_oneway(*[group['episode_0_score_for_all_state_changes'].values for
name, group in df.groupby('task_name')])

# Output the ANOVA results
anova_results
```

Programmer, please execute this Python code using the `functions.python` tool to perform the ANOVA test for Hypothesis 1. If the p-value is less than 0.05, we will reject the null hypothesis and conclude that there is a significant difference in initial performance across tasks.

Figure 12. **Background:** Experimental data of running methods using a popular agent-based repo Reflexion (`https://github.com/noahshinn/reflexion`) and (Majumder et al., 2023) was fed to DATAVOYAGER. **This figure:** Hypothesis Generation - When asked to generate hypotheses and perform sophisticated analysis based on insights (highlighted in green), the system generates testable hypotheses, formulating clear null (H0) and alternative (H1) hypotheses with corresponding statistical tests (highlighted in yellow) for uncovering complex underlying patterns in the data.

Data Modelling Analyst Expert (to chat_manager):

The distribution of task names within each cluster is as follows:

- **Cluster 0:** This cluster includes tasks such as 'find-plant', 'lifespan-longest-lived', 'find-living-thing', and 'lifespan-shortest-lived', which are the most frequent. It also contains tasks related to chemistry and genetics. These tasks have high initial and final scores, suggesting that the agent is already proficient in these areas without showing improvement, possibly because it has reached a performance ceiling.
- **Cluster 1:** This cluster is characterized by tasks that involve angles and friction ('inclined-plane-determine-angle', 'inclined-plane-friction-named-surfaces'), genetics ('mendelian-genetics-known-plant', 'mendelian-genetics-unknown-plant'), and living things ('find-living-thing', 'find-plant'). The tasks in this cluster show significant improvement, indicating that the agent is learning and adapting well to these tasks.
- **Cluster 2:** The tasks in this cluster include 'freeze', 'chemistry-mix-paint-secondary-color', and 'grow-plant'. These tasks have lower initial and final scores, with no improvement, suggesting that they are challenging for the agent and may require further investigation to understand the difficulties the agent is facing.
- **Cluster 3:** This cluster contains tasks such as 'chemistry-mix', 'grow-plant', and 'inclined-plane-friction-named-surfaces'. Despite having the lowest initial scores, tasks in this cluster show a significant average improvement. However, the final scores remain relatively low, indicating that while the agent is capable of learning, these tasks may still be too challenging or may require more complex strategies for the agent to master.

The context provided by the task names, along with the performance metrics, suggests that the agent's performance is influenced by the nature of the tasks. Tasks that are more conceptual or require understanding of complex systems (e.g., genetics, life stages) seem to be areas where the agent can learn and improve. In contrast, tasks that may involve more practical or hands-on activities (e.g., chemistry mixing, freezing) appear to be more challenging for the agent.

To further understand the agent's performance on these tasks, one could look into the specific characteristics of each task, such as the rules, objectives, and the type of learning or reasoning required. Additionally, qualitative feedback from the agent's interactions with the tasks or from experts in the domain could provide deeper insights into why certain tasks fall into specific clusters and how the agent's learning algorithm could be adjusted to improve performance.

Figure 13. **Background:** Experimental data of running methods using a popular agent-based repo Reflexion (`https://github.com/noahshinn/reflexion`) and (Majumder et al., 2023) was fed to DATAVOYAGER. **This figure:** Knowledge Frontiers Support - Ability to support and generate new insights on frontiers of knowledge, where novel insights on agents' behavior were generated. The new insights are highlighted in blue.

User Proxy (to chat_manager):

explore the data and perform different analysis to compare wealth disparities between demographics

Data Modelling Analyst Expert (to chat_manager):

To explore the data and compare wealth disparities between demographics, we can perform several analyses focusing on race and sex, as these are the demographic variables available in the dataset. Here are the analyses we can conduct:

1. **Wealth Distribution by Race and Sex**: We will examine the distribution of wealth for each race and sex category across the years 1985, 1990, and 1996.
2. **Median Wealth Comparison**: We will compare the median wealth between different race and sex groups to understand the central tendency of wealth distribution within each demographic.
3. **Wealth Inequality Measures**: We will calculate measures of wealth inequality, such as the Gini coefficient, within each demographic group to assess the level of wealth disparity.
4. **Time Series Analysis**: We will look at the changes in wealth over time for each demographic group to see if disparities are widening or narrowing.

Let's start with the first analysis, which is to examine the wealth distribution by race and sex.

Team member, please calculate the descriptive statistics for wealth distribution by race and sex for the years 1985, 1990, and 1996.

Figure 14. **Background:** National Longitudinal Survey of Youth data with a question on how incarceration and race affected wealth was fed to DataVoyager; it is a question studied in (Zaw et al., 2016). **This figure:** Knowledge Frontiers Support - Despite the original paper talking about wealth analysis post-incarceration and only doing basic statistical analysis over the data, the system was able to suggest new techniques like the application of Gini coefficients - a popular measure used in understanding wealth disparities (highlighted in green).

Data Modelling Analyst Expert (to chat_manager):

The analysis of median wealth and Gini coefficients for each demographic group across the years 1985, 1990, and 1996 provides the following insights:

Median Wealth:

- In all years, white females had the highest median wealth, followed by white males. This indicates that within the dataset, white individuals typically have more wealth than individuals of other races.
- Black individuals had the lowest median wealth across all years, with black females having the lowest median wealth in 1985 and 1990, and black males in 1996.
- Hispanic individuals had median wealth values between those of black and white individuals, with Hispanic males typically having higher median wealth than Hispanic females.

Gini Coefficients (a measure of inequality where 0 represents perfect equality and 1 represents perfect inequality):

- Black females had the highest Gini coefficient in 1985 (0.9373), indicating very high wealth inequality within this group. The Gini coefficient for black females decreased over time, reaching 0.8092 in 1996, which still indicates substantial inequality.
- Black males and Hispanic individuals also had high Gini coefficients, although they were generally lower than those for black females. The Gini coefficients for these groups decreased over time, suggesting a reduction in wealth inequality within these demographics.
- White individuals had lower Gini coefficients compared to black and Hispanic individuals, indicating less wealth inequality within these groups. The Gini coefficients for white individuals remained relatively stable over time.

Interpretation:

- The median wealth analysis shows that white individuals typically have higher wealth than black and Hispanic individuals within this dataset.
- The Gini coefficient analysis reveals that wealth inequality is highest among black females and decreases over time for all demographic groups.
- Despite the decrease in Gini coefficients, substantial wealth inequality persists within each demographic group, with black individuals experiencing the highest levels of inequality.

New Insights:

- The persistent racial wealth gap and high levels of inequality within demographic groups suggest that systemic factors continue to influence wealth accumulation and distribution.
- The reduction in Gini coefficients over time may indicate some improvement in wealth distribution within demographic groups, but the levels of inequality remain high.

New Hypotheses:

- Structural barriers such as access to education, employment opportunities, and housing may contribute to the racial wealth gap and high levels of inequality within demographic groups.
- Policies aimed at reducing inequality, such as wealth redistribution, improved access to education, and fair employment practices, may be necessary to address the disparities observed in the data.

Figure 15. **Background:** National Longitudinal Survey of Youth data with a question on how incarceration and race affected wealth was fed to DATAVOYAGER; it is a question studied in (Zaw et al., 2016). **This figure:** Hypothesis Verification - After calculating Wealth Inequality across demographic groups using gini coefficients, the system interpreted the results (highlighted in green) and generated interesting insights (highlighted in blue).

User Proxy (to chat_manager):

No, dont investigate multicollinearity issue, Instead identify any unique insights or challenges faced by different demographic groups.

Data Modelling Analyst Expert (to chat_manager):

To identify unique insights or challenges faced by different demographic groups within the context of the impact of having a criminal record on wealth accumulation, we can perform subgroup analyses. This involves examining the data for different demographic categories such as gender, race, and potentially other variables like age or education level if available.

Here are the steps we can take:

1. **Subgroup Analysis by Gender**: Compare the effects of having a criminal record on wealth between males and females.
2. **Subgroup Analysis by Race**: Examine the differences in the impact of having a criminal record on wealth among white, black, and Hispanic individuals.
3. **Interaction Effects**: Explore interaction effects between having a criminal record and demographic variables to see if the impact varies across different subgroups.
4. **Additional Demographic Variables**: If other demographic variables are available (e.g., age, education), we can include them in the analysis to uncover further insights.

Let's start with subgroup analyses by gender and race using the existing dataset. We will calculate the mean wealth for each subgroup and perform t-tests to see if there are significant differences in wealth between individuals with and without a criminal record within each subgroup.

I will assign this task to a team member to execute the Python code to perform the subgroup analyses by gender and race.

Figure 16. **Background:** National Longitudinal Survey of Youth data with a question on how incarceration and race affected wealth was fed to DATAVOYAGER; it is a question studied in (Zaw et al., 2016). **This figure:** Human Feedback Accommodation - The system performed OLS Regression, suggested the presence of multicollinearity, and removed it using Variation Inflation Factor. Then, the system was set to address multicollinearity again, but user intervention prevented redundancy and redirected it to the objective (highlighted in green).

User Proxy (to chat_manager):

yes proceed with hypothesis testing

Planner (to chat_manager):

Arguments for function call python:

```
{"cell":"from scipy import stats\n\n# Perform independent t-tests for the wealth
variables across the two groups\ntest_results_1985 =
stats.ttest_ind(df[df['ever_jailed'] == 0]['composite_wealth_1985'],\n
df[df['ever_jailed'] == 1]['composite_wealth_1985'],\n
equal_var=False)\ntest_results_1990 = stats.ttest_ind(df[df['ever_jailed'] ==
0]['composite_wealth_1990'],\n
df[df['ever_jailed'] == 1]['composite_wealth_1990'],\n
equal_var=False)\ntest_results_1996 = stats.ttest_ind(df[df['ever_jailed'] ==
0]['composite_wealth_1996'],\n
df[df['ever_jailed'] == 1]['composite_wealth_1996'],\n
equal_var=False)\n\n(test_results_1985, test_results_1990, test_results_1996)"}
```

Planner (to chat_manager):

```
(Ttest_indResult(statistic=9.779444767764458,
pvalue=6.301384191436883e-21), Ttest_indResult(statistic=9.573934679962457,
pvalue=1.555023285650679e-19), Ttest_indResult(statistic=9.721881214132456,
pvalue=7.260215807126642e-20))
```

Data Modelling Analyst Expert (to chat_manager):

The results of the independent t-tests for the wealth variables across the two groups (those with and without a criminal record) for the years 1985, 1990, and 1996 are as follows:

1. **Wealth in 1985**:
   - T-statistic: 9.7794
   - P-value: 6.3014e-21
2. **Wealth in 1990**:
   - T-statistic: 9.5739
   - P-value: 1.5550e-19
3. **Wealth in 1996**:
   - T-statistic: 9.7219
   - P-value: 7.2602e-20

Interpretation:

- In all three years, the t-statistics are positive and large, indicating that the mean wealth of individuals without a criminal record is significantly higher than that of individuals with a criminal record.
- The p-values for all three years are extremely small (far below the standard threshold of 0.05), which provides strong evidence against the null hypothesis that there is no difference in mean wealth between the two groups.

Conclusion:

- We can reject the null hypothesis for all three years and conclude that there is a statistically significant difference in wealth between individuals with and without a criminal record.
- This supports our Hypothesis 1 that individuals with a criminal record have significantly lower wealth compared to those without a criminal record.
- Given the increasing difference in mean wealth over time, there is also support for Hypothesis 2 that the wealth gap between individuals with and without a criminal record increases over time.

These results suggest that having a criminal record is associated with lower wealth accumulation, and the impact of a criminal record on wealth may become more pronounced as time progresses. This could be due to various factors, such as difficulties in finding employment, lower earning potential, or other socioeconomic barriers faced by individuals with a criminal record.

Further analysis could involve a more detailed examination of the factors contributing to this wealth gap and the long-term socioeconomic impacts of incarceration.

Figure 17. **Background:** National Longitudinal Survey of Youth data with a question on how incarceration and race affected wealth was fed to DATAVOYAGER; it is a question studied in (Zaw et al., 2016). **This figure:** Hypothesis Verification - Following the results of Descriptive Statistics, the Data Expert proposed two hypotheses. When the user prompted the system to perform Hypothesis Testing it verified them by performing T-tests, interpreted them (highlighted in green) and shared the conclusions.

Planner (to chat_manager):

Based on the given information, the objective is to generate hypotheses regarding the effect of different factors on the chances of degree completion. We will use combinations of dependent variables to model complex hypotheses.

To achieve this objective, we can follow the following plan:

1. Understand the data:
   - Analyze the columns in the dataframe to understand the variables available for analysis.
   - Identify the dependent variable, which is "BA DEGREE COMPLETED", and the predictor column, which is "SAMPLE_SEX".
2. Generate initial hypotheses:
   - Start by generating initial hypotheses based on common assumptions or prior knowledge. For example:
     - Hypothesis 1: Females are more likely to complete a BA degree compared to males.
     - Hypothesis 2: Family size has an impact on the chances of degree completion, with smaller families having higher completion rates.
     - Hypothesis 3: Higher ability scores on the ASVAB test are positively correlated with degree completion.
3. Explore combinations of dependent variables:
   - Use combinations of dependent variables to model complex hypotheses. For example, we can combine "SAMPLE_RACE" and "SES" to analyze the effect of race and socioeconomic status on degree completion.
4. Call the "run_logistic_regression" function:
   - Call the "run_logistic_regression" function with the dependent variable set as "BA DEGREE COMPLETED" and the predictor column set as "SAMPLE_SEX".
   - Analyze the results of the logistic regression model to understand the impact of gender on degree completion.
5. Repeat step 4 for other combinations of dependent variables:
   - Call the "run_logistic_regression" function with different combinations of dependent variables to analyze their impact on degree completion.
   - Analyze the results of each logistic regression model to understand the impact of the variables on degree completion.
6. Document the findings:
   - Document the findings from each logistic regression model, including coefficients, p-values, and other statistical measures.
   - Use the findings to support or reject the initial hypotheses and generate new insights regarding the effect of different factors on degree completion.
7. Seek clarity where required:
   - If there are any unclear aspects or if additional information is needed to generate hypotheses or analyze the results, seek clarification from the data modelling expert or the data provider.

By following this plan, we will be able to generate concrete hypotheses and conduct mathematical analysis to understand the effect of different factors on the chances of degree completion.

Figure 18. **Background:** Data from 1979 follow-up wave of the National Longitudinal Survey along with the question on how social background affects degree completion was fed to DATAVOYAGER; it is a research covered in (Alexander et al., 1982) **This figure:** Multi-step planning for hypothesis generation - following an initial analysis of data statistics and considering the specified goal, the system formulates a detailed plan to guide the hypothesis generation process. It generates a possible list of hypotheses (highlighted in green) and the core experimental loop (highlighted in blue.)

Data Modelling Analyst Expert (to chat_manager):

Based on the results of the various statistical tests performed, we can generate the following detailed insights and hypotheses:

1. **Interaction Insight**: The significant interaction between SES and academic ability suggests that the impact of SES on degree completion is not uniform across all levels of academic performance. This leads to the hypothesis that the benefit of higher SES on educational outcomes is moderated by a student's academic ability.
2. **Stratified Analysis Insight**: The stratified analysis indicates that the relationship between SES and degree completion varies across different levels of academic ability. In the lowest and highest quantiles of academic ability, SES has a negative coefficient, while it has a positive coefficient in the second quantile. This suggests that for students with very high or very low academic ability, SES may play a less significant role in degree completion. A possible hypothesis is that students with very high academic ability may have a high likelihood of degree completion regardless of their SES, while students with very low academic ability may face challenges in degree completion that SES cannot overcome.
3. **Mediation Analysis Insight**: The mediation analysis shows that academic ability mediates the relationship between SES and degree completion. This suggests that part of the effect of SES on degree completion operates through its impact on academic ability. A hypothesis derived from this insight is that interventions aimed at improving academic ability may reduce the disparity in degree completion rates across different SES groups.
4. **Multicollinearity Insight**: The high VIF values for SES and the interaction term suggest multicollinearity, which may be inflating the coefficients and standard errors. This indicates that the unique contributions of SES and the interaction term to the model may be difficult to disentangle. A hypothesis here is that SES and academic ability are not entirely independent predictors of degree completion, and their effects may be confounded.
5. **Model Fit and Predictive Power Insight**: The good predictive power of the model, as indicated by the AUC, suggests that the model is effective at distinguishing between individuals who complete a degree and those who do not. However, the AIC and BIC values should be used to compare this model to alternative models. A hypothesis could be that including additional relevant predictors or using alternative modeling approaches may improve the model's predictive power and fit.

From these insights, we can formulate several hypotheses for further research:

- **Hypothesis 1**: The benefit of higher SES on degree completion is greater for students with moderate levels of academic ability compared to those with very high or very low academic ability.
- **Hypothesis 2**: Academic ability is a key mediator in the relationship between SES and degree completion, and interventions that enhance academic skills could reduce SES-related disparities in educational outcomes.
- **Hypothesis 3**: The relationship between SES and degree completion is confounded by the overlap between SES and academic ability, and separating their effects requires careful modeling to account for multicollinearity.
- **Hypothesis 4**: Including additional variables related to educational support, motivation, and other socioeconomic factors may provide a more comprehensive understanding of the factors influencing degree completion.

Further research should test these hypotheses using additional data and alternative statistical methods to confirm the findings and to explore the complex dynamics between SES, academic ability, and educational outcomes.

Figure 19. **Background:** Data from 1979 follow-up wave of the National Longitudinal Survey along with the question on how social background affects degree completion was fed to DATAVOYAGER; it is a research covered in (Alexander et al., 1982) **This figure:** Hypothesis Generation - The system conducts new experiments to comprehend SES data and generates hypotheses. Building upon initial statistical tests, the model delves deeply into proposing and conducting more sophisticated experiments (highlighted in green), subsequently formulating several hypotheses for further analysis (highlighted in blue).

Data Modelling Analyst Expert (to chat_manager):

1. **Coefficients**:
   - **Family Size of Sample**: The coefficient for 'FAMILY SIZE OF SAMPLE' is positive (0.5085) and statistically significant (p < 0.001), indicating that larger family sizes are associated with a higher likelihood of completing a BA degree.
   - **Ability (Composite of ASVAB Score)**: The coefficient for 'ABILITY: COMPOSITE OF ASVAB SCORE' is negative (-0.0021) and statistically significant (p < 0.001), suggesting that higher ability scores are associated with a slightly lower likelihood of completing a BA degree, which is counterintuitive and may require further investigation to understand the context.
   - **Percentile in Class**: The coefficient for 'PERCENTILE IN CLASS' is negative (-0.0258) and statistically significant (p < 0.001), indicating that higher class rankings are associated with a lower likelihood of completing a BA degree, which is also counterintuitive.
   - **SES (Socioeconomic Status)**: The coefficient for 'SES' is negative (-0.2750) and statistically significant (p = 0.014), suggesting that higher socioeconomic status is associated with a lower likelihood of completing a BA degree, which may seem surprising and warrants further exploration.
2. **Model Fit**:
   - The Pseudo R-squared value is 0.2246, which indicates that the model explains approximately 22.46% of the variance in BA degree completion based on the included predictor variables.
   - The model has converged, and the LLR (Likelihood Ratio Test) p-value is highly significant (p < 0.001), indicating that the model as a whole fits significantly better than an empty model (one with no predictors).

These results suggest that there are significant associations between the predictor variables and BA degree completion. However, the direction of some of these associations is not as expected. It is important to consider that these results are based on the data provided and the model specified. There may be other factors not included in the model that could explain these findings, or there may be interactions between variables that are not accounted for.

Further investigation into the context of the data, potential confounding variables, and the possibility of non-linear relationships or interactions between variables would be necessary to fully understand the dynamics of BA degree completion in this sample.

Figure 20. **Background:** We inverted the status of graduation in the 1979 follow-up wave of the National Longitudinal Survey dataset to verify the robustness of the DATAVOYAGER. We asked the usual query: how social background affects degree completion to DATAVOYAGER; **This figure:** The system correctly detects and communicates the surprising result (inverted trend) to the user (highlighted in red). An ideal data-discovery system should have the ability to detect and flag surprises in data.