Hypothesis Exploration with LLM-driven Intelligent Orchestration System

Bruno G. Tavares dos Santos
UFRGS
Porto Alegre, Brazil
bruno.tsantos@inf.ufrgs.br

Vicente Nejar de Almeida
Instituto de Informática, UFRGS
Porto Alegre, Brazil
vicente.nejar@gmail.com

Eduardo Ribeiro
University of Tocantins (UFT)
Palmas, Brazil
uft.eduardo@uft.edu.br

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alps
Grenoble, France
sihem.amer-yahia@univ-grenoble-alpes.fr

João Luiz Dihl Comba
UFRGS
Porto Alegre, Brazil
comba@inf.ufrgs.br

ABSTRACT

Exploratory data analysis over complex datasets requires analysts to formulate and test a large number of statistical hypotheses, a process that becomes increasingly difficult as the combinatorial search space grows. While traditional visual analytics systems can support exploration, they still rely heavily on manual hypothesis formulation. Conversely, naive applications of Large Language Models (LLMs) to raw tabular data often produce superficial insights or hallucinated conclusions due to the lack of statistical grounding. We present **Hypothesis Exploration with LLM-driven Intelligent Orchestration System (HELIOS)**, an interactive, multi-agent visual analytics system that combines automated hypothesis discovery with LLM-based analytical synthesis. Candidate hypotheses are statistically evaluated and ranked according to coverage, diversity, impact, and homogeneity, while specialized agents synthesize the results into interpretable narratives and suggest further analytical directions. In this demonstration, users can upload their own datasets or use available sample datasets to explore relationships among demographic, contextual, and behavioral attributes. This setting allows analysts to investigate which attribute combinations drive higher values of a target metric for specific subgroups, how contextual or temporal factors influence outcomes, and which subpopulations exhibit distinct or unexpected patterns.

PVLDB Reference Format:
Bruno G. Tavares dos Santos, Vicente Nejar de Almeida, Eduardo Ribeiro, Sihem Amer-Yahia, and João Luiz Dihl Comba. Hypothesis Exploration with LLM-driven Intelligent Orchestration System. PVLDB, 19(12): 4814 - 4817, 2026.
doi:10.14778/3827998.3828129

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/brunogtds/helios.

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3828129

1 INTRODUCTION

Exploratory data analysis based on multiple hypothesis testing has become an important approach for extracting insights from complex datasets, often requiring analysts to formulate and validate a large number of statistical hypotheses [3, 5, 6]. This process is inherently challenging, as the combinatorial search space of possible subgroups grows rapidly with the number of attributes.

Traditional visual analytics tools support manual exploration [4], but analysts must still identify promising hypotheses themselves, which limits scalability and may introduce cognitive bias. Prior systems, such as SHEVA [1], address this challenge by automatically generating statistically sound hypotheses using reinforcement learning and optimizing criteria such as coverage and diversity. While effective in identifying valid hypotheses, these approaches can be computationally expensive, particularly when training from scratch, and offer limited support for interpreting results.

Large Language Models (LLMs) have emerged as accessible tools for supporting data analysis tasks [7], enabling natural language interaction, summarization, and explanation. However, LLMs alone are not well-suited for statistical discovery, as they lack rigorous mechanisms for conditional hypothesis testing and may rely on superficial correlations or produce unsupported conclusions.

To address these limitations, we present **HELIOS**, an interactive system that extends prior work [1] through an orchestrated multi-agent architecture that integrates automated hypothesis discovery with analytical synthesis. Rather than relying on a single model, **HELIOS** coordinates specialized agents that collaboratively explore the hypothesis space, guided by user intent and dataset context. The Hypothesis Agent generates candidate subgroups, which are subsequently validated through statistical tests (e.g., t-tests, chi-square tests, and Kolmogorov–Smirnov tests) to ensure significance before being passed to agents for interpretation and further analysis.

The key contribution of **HELIOS** lies in a coordinated, human-in-the-loop multi-agent workflow, in which agents assume roles in exploration, validation, and interpretation. This orchestration enables systematic navigation of large hypothesis spaces while preserving statistical rigor and improving interpretability. We demonstrate **HELIOS** using benchmark and user-uploaded datasets, as well as real-world data, highlighting its ability to uncover meaningful, non-trivial patterns across diverse analytical scenarios.

2 THE HELIOS SYSTEM

HELIOS is a multi-agent system that combines automated statistical discovery with LLMs to support interactive, hypothesis-driven data analysis. As illustrated in Figure 1, the architecture is organized into three main components: *data preparation*, *hypothesis generation and validation*, and *LLM-driven narrative synthesis*. Together, these components enable analysts to move from raw tabular data to interpretable insights while preserving statistical rigor.

2.1 Data Preparation

The first stage prepares the dataset for hypothesis discovery. The user uploads a dataset and specifies a target variable representing the outcome of interest. Numerical attributes are optionally discretized to enable subgroup-based comparisons.

During this phase, the system performs automatic data profiling to extract descriptive statistics and structural properties of the dataset, including attribute distributions (e.g., mean, median, mode), frequent values, missingness, and candidate categorical partitions. These summaries provide contextual information that guides both hypothesis generation and subsequent interpretation. The user may also specify an exploratory query (e.g., *"What patterns exist across different regions?"*), which steers the search toward relevant data regions. HELIOS does not impute missing values. Each test uses complete cases for its target and subgroup attributes, while coverage remains relative to the original dataset. The profile reports missingness, and candidates with insufficient observations are rejected with a warning; hence, severely incomplete data are not treated as complete. The output of this stage is a structured representation of the dataset enriched with statistical summaries and user intent, which is passed to the hypotheses generation module.

2.2 Hypothesis Generation and Validation

The second stage explores the hypothesis space and identifies statistically significant hypotheses. The **Hypothesis Agent** receives (1) a data context, which comprises a semantic description of the dataset and global statistics for all attributes, and (2) a user-defined query that provides guidance for focusing on specific goals.

Given this data context and the selected target variable, the system generates hypotheses aligned with the user's specified goals. Each hypothesis represents an assertion about the value or distribution of the target variable within a data region, defined as a subgroup characterized by a combination of attribute conditions (e.g., `occupation = 'Programmer' AND age = '18-24'`). Appropriate statistical tests are applied to evaluate each hypothesis and determine whether the observed deviation from the global baseline is statistically significant. Only hypotheses that satisfy the significance criteria are retained and presented to the user. The query is a soft constraint: named attributes are prioritized, whereas an open-ended prompt such as *"find interesting patterns"* triggers exploration across all eligible attributes using the data profile. Both modes undergo identical statistical validation and ranking.

The resulting hypotheses are ranked using multiple complementary criteria. Let D denote the dataset and $D_i \subseteq D$ the subset covered by hypothesis h_i. Each hypothesis is evaluated according to: (i) *q-value*, controlling false discoveries via FDR [2]; (ii) *coverage*, defined as $|D_i|/|D|$; (iii) *diversity*, measured via pairwise Jaccard distance; (iv) *impact*, defined as $\mu(D_i)/\mu(D)$; and (v) *homogeneity*, defined as $1/\sigma^2(D_i)$. The q-value serves only as an eligibility filter, discarding candidates above the configured FDR threshold. The remaining candidates are ranked by the min–max normalized combination of coverage, diversity, impact, and homogeneity as $S(h_i) = \frac{1}{4}(\widehat{C}_i + \widehat{D}_i + \widehat{I}_i + \widehat{H}_i)$. The demo uses equal weights by default but supports configurable weighting schemes.

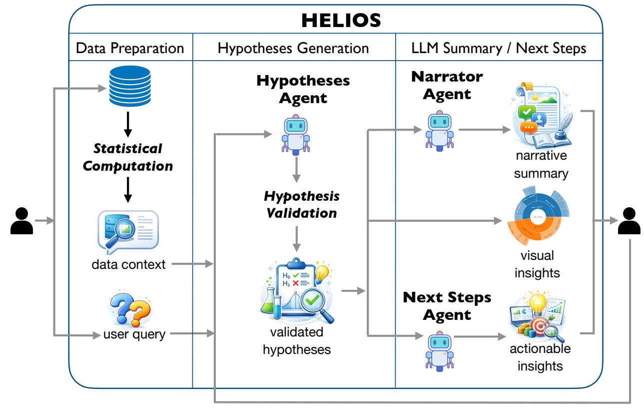

Figure 1: HELIOS System architecture.

2.3 LLM-driven Summary and Next-Steps

The final stage presents the validated hypotheses and supports an iterative analysis process, bridging statistical discovery and human interpretation by transforming results into actionable insights. Hypotheses are displayed through an interactive dashboard that offers multiple perspectives, including a structured table of statistical quality metrics (e.g., q-value, coverage, and impact), a collapsible Tree View for inspecting condition hierarchies, and visualizations such as Sunburst charts for high-level analysis.

Interpretation is provided by two LLM agents. The **Narrator Agent** generates coherent analytical narratives that summarize key findings and highlight their implications, optionally incorporating external knowledge to provide plausible explanations for observed patterns. The **Next-Steps Agent** analyzes the current results and suggests directions for further exploration, guided by user intent. These suggestions may include refining subgroup definitions, exploring alternative perspectives, or generating new hypotheses.

Narratives separate validated results from conjectural explanations: numerical claims come from validated records, and any suggested subgroup must be tested before becoming a finding. This reduces, but does not eliminate, hallucination risk.

All insights and recommendations are presented to the user, who can iteratively refine the analysis by updating the intent and exploring new hypotheses. The orchestration is sequential: predicates are validated and ranked by a non-LLM module before reaching the Narrator and Next-Steps Agents, preventing LLMs from modifying statistical results or reinstating rejected hypotheses.

3 HELIOS USER INTERFACE

The **HELIOS** interface is designed to support the full analysis workflow through a single-page web application that emphasizes

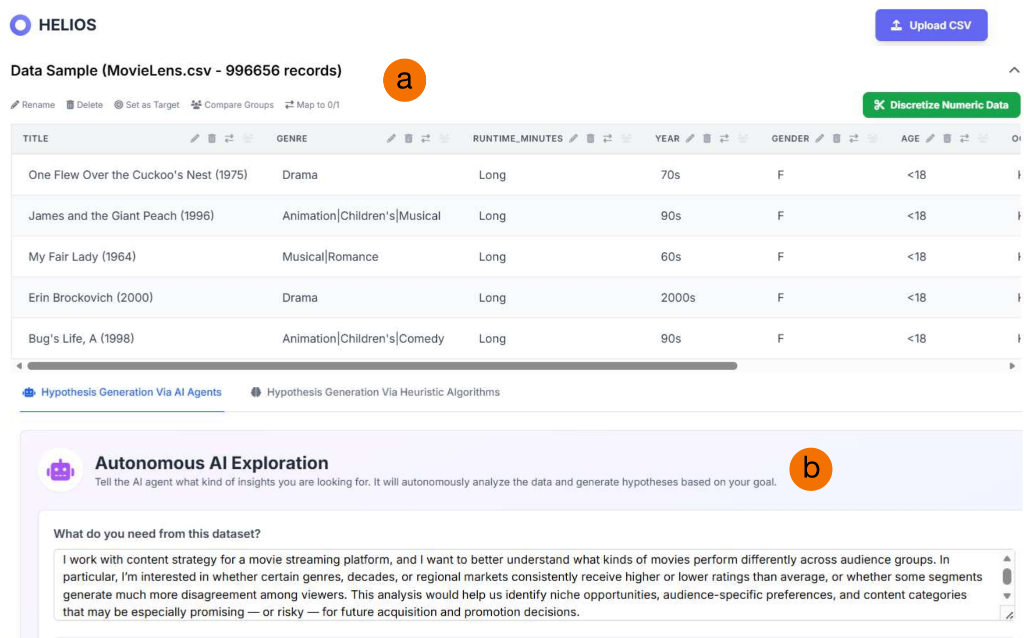

Figure 2: HELIOS user interface: (a) loading a dataset for analysis and (b) entering a user query.

usability and interactive exploration. Rather than exposing the underlying system components, the interface centers the interaction around a natural language query, guiding users from data input to insight generation in a structured yet flexible manner. Users need not know which attributes matter. After selecting an outcome, HELIOS produces a cross-attribute overview with ranked evidence and next-step suggestions for iterative refinement.

Users begin by uploading a dataset and selecting a target variable. A key element of the interface is the ability to express analytical intent as a natural-language query (e.g., identifying factors associated with a target variable or characterizing specific subgroups), which directly drives the hypothesis exploration process (Figure 2). Users also select the model and output size; temperature appears only for models that support it. Together, these controls allow users to steer the analysis while maintaining flexibility across different use cases.

Once the analysis is executed, results are presented through coordinated visualizations that support comparison, filtering, and navigation. The main *Table View* displays hypotheses ranked by their overall score, allowing users to quickly identify the most relevant patterns. Complementary views, such as the *Tree View* and *Sunburst View*, provide hierarchical perspectives on the hypothesis space, helping users understand how subgroups are constructed and how different attribute combinations relate to one another.

The interface further supports iterative exploration by integrating system-generated narratives and recommendations directly into the workflow. Users can inspect explanations, explore suggested directions, and refine their analytical intent without leaving the environment. This tight integration between interaction, visualization, and guidance enables users to continuously explore, validate, and extend their analysis in a cohesive manner.

Figure 3 provides an overview of the interface, combining tabular and hierarchical visualizations of hypotheses with agent-generated summaries and recommendations for continued exploration.

4 CASE STUDY

We present a case study evaluating **HELIOS** using a subset of the Stanford Open Policing Project dataset comprising 86,113 traffic stop records from Rhode Island, with the binary arrest indicator as the target variable. The system is configured to support a civil rights audit by jointly optimizing statistical significance, impact, coverage, diversity, and subgroup homogeneity. The Hypothesis Agent explores three complementary strategies: identifying subgroups with high arrest likelihood (procedural precursors), low arrest likelihood (privilege profiles), and high outcome variability (discretionary inconsistency). All hypotheses are validated using statistical tests for differences in means and variances, with false discovery rate control to ensure robustness.

The system identifies several meaningful patterns while filtering out tautological artifacts such as `stop_outcome = 'Arrest Driver'`. Among the non-trivial findings, Black drivers held for 16–30 minutes for an *Other Traffic Violation* exhibit a 1.41× higher arrest rate (coverage: 2.9%, $q = 0.0153$), highlighting a significant demographic interaction. In contrast, after removing confounding factors, older driver cohorts (`driver_age_raw_class = '1987-8801'`) searched for equipment violations show a protective effect with a 0.65× impact lift (coverage: 7.3%, $q < 0.0001$). The system also uncovers patterns of discretionary inconsistency: Black drivers held for 16–30 minutes without drug involvement present significantly higher outcome variance (impact lift: 1.11×, coverage: 8.0%, $q < 0.0001$), suggesting systemic variability rather than isolated anomalies. These results demonstrate **HELIOS** 's ability

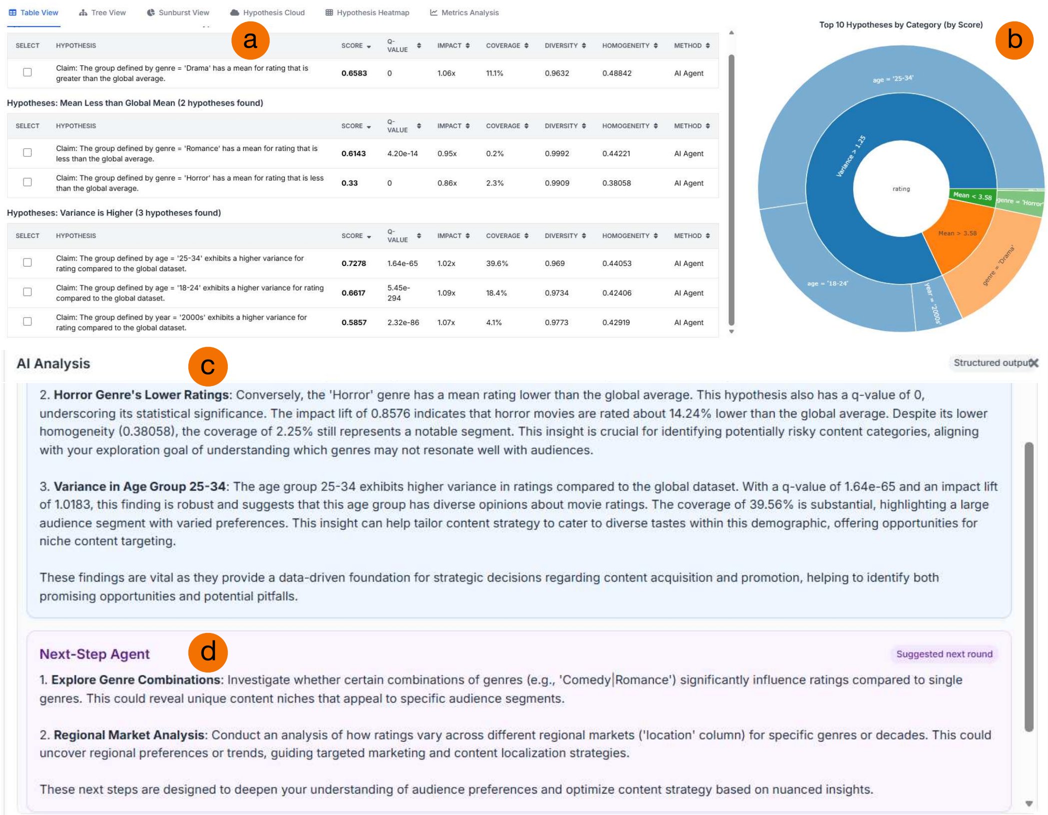

Figure 3: HELIOS displays a table (a) and hierarchy (b) of significant hypotheses. The Narrator agent (c) summarizes the results, while the Next-Step agent (d) suggests continuations of the exploration process.

to navigate large combinatorial spaces and uncover statistically robust, socially meaningful subgroup patterns that are difficult to identify through manual analysis. This case study demonstrates feasibility, not superiority; controlled comparisons of discovery quality, runtime, and narrative factuality remain future work.

5 DEMONSTRATION SCENARIO

In our demonstration, attendees use a provided dataset (e.g., MovieLens) or upload their own, select a target variable, optionally configure preprocessing (e.g., binning numerical attributes), and specify a high-level analytical intent. They may also begin with a vague prompt, relying on **HELIOS** to identify relevant explanatory attributes. A statistical profiling module computes descriptive summaries, distributions, missingness indicators, and other data quality statistics that guide the **Hypothesis Agent** in proposing statistically testable subgroups. A non-LLM module evaluates these candidates using appropriate statistical tests, applies multiple-comparison correction, and retains only validated and diverse hypotheses. Results are presented through coordinated Table, Tree, and Sunburst views with statistical quality metrics. The **Narrator Agent** summarizes validated findings, while the **Next-Steps Agent** recommends refinements, alternative perspectives, or new hypotheses. Users can inspect the evidence, refine their intent, compare successive iterations, and export the resulting analysis.

ACKNOWLEDGMENTS

Partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] Vicente Almeida, Eduardo Ribeiro, Nassim Bouarour, João Comba, and Sihem Amer-Yahia. 2023. SHEVA: A Visual Analytics System for Statistical Hypothesis Exploration. *Proceedings of the VLDB Endowment* 16 (09 2023), 4102–4105.

[2] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29 (2001), 1165–1188.

[3] Nassim Bouarour, Idir Benouaret, and Sihem Amer-Yahia. 2022. *Optimizing Data Coverage and Significance in Multiple Hypothesis Testing on User Groups.* Springer Berlin Heidelberg, Berlin, Heidelberg, 64–96.

[4] Wenqiang Cui. 2019. Visual Analytics: A Comprehensive Overview. *IEEE Access* 7 (2019), 81555–81573. https://doi.org/10.1109/ACCESS.2019.2923736

[5] Doris Jung-Lin Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya Parameswaran. 2019. Avoiding drill-down fallacies with VisPilot: assisted exploration of data subsets. In *Proc. of the 24th International Conference on Intelligent User Interfaces*.

[6] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, and Srividya Subramanian. 2021. DORA THE EXPLORER: Exploring Very Large Data With Interactive Deep Reinforcement Learning. In *Conference on Information and Knowledge Management*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 4769–4773.

[7] Jiani Zhang, Hengrui Zhang, Rishav Chakravarti, Yiqun Hu, Patrick Ng, Asterios Katsifodimos, Huzefa Rangwala, George Karypis, and Alon Halevy. 2025. CoddLLM: Empowering Large Language Models for Data Analytics. arXiv:2502.00329 [cs.DB]